Journal of Korean Library and Information Science Society
Vol. 57, No. 3

OPEN ACCESS

국가서지 LOD의 시맨틱 검색 가능성 모색 연구

- 한국 현대 소설 분야의 도서를 중심으로 -

A Study on the Feasibility of Semantic Search in the National Library of Korea Linked Open Data: Focusing on Books in Modern Korean Fiction

송 민 건(Min-geon Song)*
이 수 상(Soo-Sang Lee)**

〈 목 차 〉

Ⅰ. 서 론
Ⅱ. 이론적 배경
Ⅲ. 국가서지 LOD의 데이터 구조 분석
Ⅳ. 시맨틱 검색의 적용 및 개선
Ⅴ. 논의 및 결론

요 약 : 국립중앙도서관 국가서지 LOD는 의미적 연결과 활용 확대를 목적으로 하나, 그 구조가 시맨틱 검색을 어느 수준까지 지원하는지는 확인되지 않았다. 이에 본 연구는 국가서지 LOD 데이터 구조의 특징을 분석하고 그 구조에서 시맨틱 검색이 수행되는 수준을 파악하였다. 링크드 데이터의 발행 원칙에서 발행과 활용의 요건 4가지를 정리한 뒤, 국가서지 LOD 데이터셋에서 한국 현대 소설로 분류된 도서와 이에 연결된 저자, 주제를 트리플 저장소에 적재해 SPARQL 질의로 분석하였다. 분석 결과는 다음과 같다. 첫째, 온톨로지가 클래스 31종을 선언하였으나 개체를 실질적으로 구분하는 것은 도서, 저자, 주제의 3가지였다. 둘째, 개체를 연결하는 관계는 4가지 유형이었고, 도서와 저자는 단일 속성으로만 이어져 역할이 구분되지 않았다. 셋째, 값과 연결이 채워지지 않아 저자 이름을 가진 도서의 32.95%에 저자 URI 연결이 없었고 역방향 트리플 18,827건이 비어 있었다. 시맨틱 검색은 데이터에 적힌 관계를 따라가는 데 머물렀고, 개체와 관계의 풍부함이라는 조건은 아직 갖추어지지 않았다. 클래스 신설, 관계의 세분화, 값과 연결의 정비를 개선 방안으로 제안하였다.

주제어 : 국가서지 LOD, 링크드 오픈 데이터, 시맨틱 검색, 온톨로지, SPARQL

ABSTRACT : The National Library of Korea publishes its national bibliography as the National Library of Korea Linked Open Data with the aim of semantic connection, yet the extent to which that structure supports semantic search has not been established. This study analyzed the characteristics of the data structure and examined the level at which semantic search is performed on it. Four requirements for publication and use were derived from the linked data principles, and books classified as modern Korean fiction in the National Library of Korea Linked Open Data datasets, together with the authors and subjects linked to them, were loaded into a triple store and analyzed with SPARQL queries. The findings are as follows. First, although the ontology declared 31 classes, the entities were effectively organized by three: books, authors, and subjects. Second, four types of relations connected the entities, and books and authors were linked by a single property, so roles were not distinguished. Third, values and links were left unfilled: 32.95% of the books carrying an author name string had no link to an author URI, and 18,827 inverse triples were missing. Semantic search remained confined to the relations recorded in the data, and the richness of entities and relations on which it depends was not yet in place. Declaring new classes, subdividing relations by role, and maintaining values and links were proposed as directions for improvement.

KEYWORDS : National Bibliography, Linked Open Data, Semantic Search, Ontology, SPARQL

* 부산대학교 문헌정보학과 박사수료(mgs207@pusan.ac.kr / ISNI 0000 0005 1420 3658) (제1저자)
** 부산대학교 문헌정보학과 교수(sslee@pusan.ac.kr / ISNI 0000 0000 6434 9851) (교신저자)

• 논문접수: 2026년 8월 17일 • 최초심사: 2026년 9월 4일 • 게재확정: 2026년 9월 18일
• 한국도서관·정보학회지, 57(3), 311-340, 2026. http://dx.doi.org/10.16981/kliss.57.3.202609.311

※ Copyright © 2026 Korean Library and Information Science Society
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

Ⅰ. 서 론

1. 연구의 배경과 목적

국립중앙도서관은 「도서관법」 제20조 제1항 제3호에 따라 국가 서지(書誌) 작성 및 표준화를 업무로 한다. 이에 따라, 국립중앙도서관은 국내에서 간행된 자료의 서지정보를 종합적으로 수록한 대한민국 국가서지를 발간하며, 이 국가서지를 RDF 데이터로 변환한 국가서지 LOD(Linked Open Data)를 발행하고 있다.

국가서지 LOD 발행 목적은 데이터 공개에 그치지 않는다. 국립중앙도서관(2013)은 HTML 문서 위주의 연결성을 벗어나 원천 데이터를 표준 형식에 맞게 개방하여 데이터 중심의 의미적 연결성을 확보하는 것을 목적으로 밝혔다. 국립중앙도서관(2024)은 또한 '웹 환경에 적합한 서지 데이터 구축 및 제공, 새로운 자원 유형에 대한 원활한 수용, 도서관 내·외부 데이터 간 상호 재활용성 제고'를 목적으로 들면서, 2030년을 목표로 국가서지 데이터 체계를 새로운 서지 구조로 전환할 것을 준비하고 있다고 서술하였다. 국가도서관위원회(2024)의 제4차 도서관발전종합계획도 '도서관 데이터의 관리·활용 체계 확립'을 과제로 두었고, 2026년도 시행계획(국가도서관위원회, 2026)에서 국립중앙도서관은 '국가 지식자원 데이터를 의미적으로 연결함으로써 데이터의 확장성 및 활용성 제고'를 사업 목표로 제시하였다. 4개 문헌 모두 국가서지 LOD 데이터의 의미적 연결과 활용 확대를 발행 목적으로 삼고 있다.

Berners-Lee(2006)는 링크드 데이터의 원칙을 지키지 않으면 데이터가 서로 연결될 기회를 잃고 뒤에 예기치 않게 재사용될 길이 좁아진다고 하면서, 웹이 더하는 가치가 정보의 예기치 않은 재사용에 있다고 하였다. 데이터를 내려받아 다른 환경에서 다시 쓰는 것이 여기에 해당한다. 국가서지 LOD 홈페이지(https://lod.nl.go.kr)는 검색 서비스와 데이터셋 다운로드, OpenAPI와 SPARQL Endpoint를 제공한다.

본 연구에서는 현재 국가서지 LOD의 데이터 구조에서 관계 기반 탐색과 추론을 통한 시맨틱 검색의 가능성을 확인하고자 한다. 그러나 국가서지 LOD가 시맨틱 검색을 어느 수준까지 지원하는지는 기관의 공식 소개나 계획의 서술만으로는 알 수 없다. 따라서 현재의 데이터 구조에서 시맨틱 검색의 수준을 파악할 필요가 있다. 이에 본 연구는 국가서지 LOD의 데이터 구조를 분석하고, 그 구조 위에서 시맨틱 검색이 어느 수준까지 수행될 수 있는지를 확인하여, 그 수준을 높이기 위한 구조 개선의 방향을 제시하고자 하였다.

분석의 대상은 국가서지 LOD가 다운로드 형식으로 제공하는 데이터셋이며, 그 가운데 일반도서 데이터셋의 레코드 약 388만 건 가운데 오프라인 자료의 서지레코드로서 한국 현대 소설(KDC 813.7)로 분류된 도서 53,548건을 사례로 삼아 분석하였다. 해당 사례를 선정한 이유는 한국 현대 소설이라는 하나의 분류기호 안에서도 장르와 시대 배경이 여럿으로 나뉘고, 글과 그림, 번역처럼 저자의 역할도 여러 가지로 나타나므로 이러한 구분을 데이터가 개체와 관계로 표현하고 있는지

를 확인하기에 적절하다고 보았기 때문이다.

본 연구는 다음과 같은 과정을 거쳤다. 첫째, 데이터셋 파일을 다운로드하여 사례에 해당하는 도서를 추출하고, 그 도서에 연결된 개체를 파악하여 함께 추출한 뒤 트리플 저장소에 적재하였다. 둘째, 적재한 사례 데이터에서 데이터 구조의 특징을 도서 한 건의 개체 구성에서 개체별 속성의 구성, 전체 그래프의 규모, 온톨로지와 관계의 구성까지 차례로 확인하였다. 셋째, 같은 저장소에서 명시된 관계를 질의로 따라가는 탐색과 선언으로부터 관계를 도출하는 추론을 수행하여 그 제한점을 확인하였다. 넷째, 셋째 단계에서 드러난 제한에 대응하는 개선 방안 가운데 온톨로지 선언의 보완을 적용한 데이터를 따로 구성하여, 탐색과 추론이 어디까지 이르는지를 다시 확인하였다.

본 연구에서 설정한 연구문제는 다음과 같다.

연구문제 1. 국가서지 LOD의 데이터 구조는 어떤 특징을 갖는가?

연구문제 2. 국가서지 LOD의 현행 데이터 구조에서 시맨틱 검색은 어느 수준까지 수행될 수 있는가?

연구문제 3. 국가서지 LOD에서 시맨틱 검색 수준을 높이려면 어떤 구조 개선이 가능한가?

2. 선행연구

본 연구와 관련된 선행연구는 국가서지 LOD를 대상으로 한 연구와 검색과 주제 접근을 다룬 연구의 2가지 측면으로 나누어 살펴볼 수 있다.

먼저 국가서지 LOD의 구축과 발행을 다룬 연구는 다음과 같다. 조명대(2010)와 이성숙(2020)은 해외 사례와 국내 구축 현황을 검토하여 링크드 데이터의 적용 모델과 개선 방안을 제시하였고, 이미화와 이지원(2021)은 주요국의 국가서지를 조사하여 생성 방식의 변화가 모색되어야 한다고 보았다. 노지현(2019)은 편목의 관점에서 국내 3개 기관이 구축한 링크드 데이터를 검토하여, 기존 데이터를 링크드 데이터 형식으로 변환하여 공개한 것 이상의 의미를 발견하기 어려우며 링크드 데이터에서 핵심이라 할 수 있는 데이터 사이의 연결이 거의 반영되어 있지 못하다고 보았다. 이러한 연구들의 관심은 국가서지를 링크드 데이터로 어떻게 구축하고 발행할 것인가에 있다. 노지현(2019)이 데이터 사이의 연결을 지적한 것처럼 발행된 결과에 대한 평가도 있으나, 그 데이터로 어떤 검색이 가능한지를 확인한 것은 아니다.

발행된 데이터 자체를 대상으로 한 연구도 수행되었다. Deliot et al.(2016)은 영국 국가서지 링크드 오픈 데이터의 이용을 분석하여, 접근의 양상이 기계 접근이 다수를 차지하는 쪽으로 옮겨 갔음을 확인하였다. 박진호(2026)는 2026년 4월 1일자 국가서지 LOD 전체를 SHACL과 ShEx로 검증하여, 온톨로지 설계에서 전체 속성의 41.4%가 정의역을 명시하지 않는 명세의 불완전성을 확인하였다. 이들 연구는 발행된 데이터를 대상으로 삼는다는 점에서 본 연구와 대상이 같다. 다만 Deliot et al.(2016)이 본 것은 그 데이터를 누가 어떻게 이용하는가이고, 박진호(2026)가 본 것은 개별 자원의

기술이 온톨로지 설계 명세를 지키는가이다. 후자의 검증은 필수 속성과 값의 형식을 판정할 뿐, 속성의 값이 다른 자원을 가리키는지와 그 값을 따라 탐색을 수행할 수 있는지는 판정하지 않는다.

다음으로 주제 접근이 검색을 어떻게 뒷받침하는지를 다룬 연구는 다음과 같다. 백지원과 정연경(2014)은 국립중앙도서관 주제명표목표 검색 시스템의 문제점을 파악해 개선안을 제안하였고, 최윤경과 정연경(2014)은 주제명 기술과 계층 관계 적용의 비일관성을 문제점으로 확인하였다. 노지현(2015)은 대한민국 국가서지의 번역서 주제명을 분석하여, 소설에 부여된 주제명이 대부분 문학형식을 나타내는 데 그쳐 주제 검색의 기초 데이터로 삼기에 한계가 있다고 보았다. 이혜경과 이용구(2023)는 국가서지에 부여된 주제명을 분석하여, 이용자의 탐색을 지원하려면 파편화되어 있는 주제명 사이의 계층 관계를 재정립할 필요가 있다고 보았다. 정도헌(2024)은 시맨틱 추론 엔진을 개발하여 주제명표목표의 논리적 오류를 검출하였다. 다만 여기에서의 추론은 이미 설정된 관계에서 오류를 찾아내는 것으로, 선언으로부터 관계를 도출하는 것과는 다르다. 이러한 연구들이 살펴본 것은 포털의 검색 화면(백지원, 정연경, 2014)과 주제명표목표(정도헌, 2024; 최윤경, 정연경, 2014), 그리고 MARC 형식의 서지레코드에 부여된 주제명(노지현, 2015; 이혜경, 이용구, 2023)이다.

해외에서는 링크드 데이터를 활용한 검색에 관한 연구가 수행된 바 있다. Aalberg et al.(2019)은 유럽 국가도서관 목록의 MARC 레코드를 BIBFRAME과 LRM 2가지 모델로 변환하여, 모델에 따라 생성되는 개체와 관계의 수를 비교하고 키워드 검색을 구현하였다. 이러한 연구들이 다룬 것은 검색을 위하여 새로 만든 데이터와 시스템이며, 이미 발행되어 있는 데이터가 아니다.

발행된 국가서지 LOD에 질의를 수행하여 데이터 구조가 어떤 특징을 가지며 그 위에서 탐색과 추론이 어디까지 이르는지를 다룬 연구는 확인되지 않는다. 본 연구는 발행된 국가서지 LOD를 SPARQL 질의로 직접 분석하여 데이터 구조의 특징과 시맨틱 검색의 수준을 확인하고, 이를 높이기 위한 데이터 구조 개선의 방향을 제시하였다는 점에서 선행연구와 차별성을 가진다.

Ⅱ. 이론적 배경

1. 국가서지 LOD와 링크드 데이터

국립중앙도서관은 데이터 중심의 의미적 연결성 확보를 목적으로 국가서지 LOD를 발행하고 있다. 현재 국가서지 LOD는 서지데이터와 전거데이터를 데이터셋 단위로 제공한다. 전거데이터는 개인명, 단체명, 주제명으로 나뉜다.

국가서지 LOD와 같은 링크드 데이터는 RDF 형식으로 표현될 수 있다. RDF는 자원을 URI로 식별하고, 한 자원을 다른 자원이나 값에 잇는 진술을 단위 데이터로 삼는다. 이 진술은 주체(subject),

속성(predicate), 객체(object)의 3가지 자리로 이루어지며 이를 트리플(triple)이라 한다. 주체는 URI로 식별되는 자원이고, 객체는 주체를 서술하는 속성의 값이다. 객체는 URI로 식별되는 다른 자원이거나 리터럴(literal)로 표시된다. 리터럴은 문자열이나 숫자 같은 값이다. 객체가 URI 참조이면 그 속성은 주체와 객체를 잇는 관계 속성이 되고, 객체가 리터럴이면 주체 자원의 리터럴 속성이 된다.

월드 와이드 웹 컨소시엄(World Wide Web Consortium, W3C)에 따르면 트리플은 주체와 객체를 연결하는 속성의 단위 그래프이며, 이들이 모여 이루는 집합을 RDF 그래프라 한다(W3C, 2014). 그래프에서 일부 트리플을 떼어 낸 부분집합도 그 자체로 그래프를 이루며, 이를 서브그래프(subgraph)라 한다. 트리플 하나도 하나의 서브그래프이며, 한 자원에 딸린 트리플들이나 질의 조건과 일치하는 트리플들이 각각 서브그래프가 된다. 곧 RDF 데이터는 표를 행과 열로 나누는 관계형 데이터가 아니라 개체와 개체를 잇는 그래프형 데이터이며, 이러한 데이터를 그래프 구조로 정리하여 관리하는 것이 그래프 데이터베이스이다. 자원의 유형(범주)은 클래스로 묶이고, 속성과 클래스는 함께 하나의 어휘를 이루며, 각 어휘는 고유한 네임스페이스로 식별된다.

Berners-Lee(2006)는 링크드 데이터의 발행 원칙 가운데 넷째로 다른 URI로의 링크를 포함할 것을 들었다. 이때 연결을 실제로 이루는 것이 RDF 링크이다(Bizer et al., 2009). 트리플이 주체와 객체를 잇는 연결로 기능하려면 객체가 URI 참조를 통해 식별될 수 있어야 한다. 그러므로 시맨틱 검색을 지원하기 위해서는 개체들이 URI 참조로 관리되어야 한다.

W3C(2011)는 도서관 데이터를 링크드 데이터로 발행하면 도서관과 도서관 밖의 정보 자원을 가로지르는 항해(navigation)가 더 정교해질 것이라고 하였다. 이 항해가 얼마나 정교하게 이루어지는지가 시맨틱 검색의 수준이 되며, 그것은 발행된 데이터의 구조에 달려 있다. 본 연구는 이 구조 위에서 국가서지 LOD의 탐색과 추론이 어디까지 성립하는지를 확인한다.

2. 검색 개념과 시맨틱 검색

검색은 기본적으로 이용자의 질의를 데이터베이스와 대조하여 결과를 반환하는 활동이다. 정보검색(information retrieval)이라는 이름으로 연구되었으며, 본 연구에서는 검색으로 통칭한다. 도서관 목록에서 검색의 원형은 조회(lookup)이다.

검색의 대상이 되는 데이터베이스가 늘면서 여러 데이터베이스를 한 번에 검색하려는 요구가 나타났고(Diao, 2018), 이는 메타데이터를 모아 통합 색인 위에서 검색하는 디스커버리(discovery)로 구현되었다.

검색의 대상은 이미 알고 있는 자료를 찾는 조회에서, 구체적으로 표현되지 않은 요구를 따라가는 탐색으로, 나아가 아직 요구로 정식화되지 않은 자원과 만나는 세렌디피티(serendipity)로 확대되고 있다. 우연한 발견 자체는 특정한 질의 없이 자료를 훑어보는 브라우징과 디스커버리 환경에서도 나타난다. 세렌디피티는 그 가운데 자원의 유형과 관계의 의미가 미리 정의된 온톨로지(ontology)

기반 환경에서 일어나는 우연한 발견을 가리킬 때 주로 사용한다.

Gnoli et al.(2024)은 시소러스의 관련어를 이용하여 관련 주제를 둘러보거나 하위어와 상위어를 보는 일이 가능하지 않으면 탐험(explore)이 성립하지 않는다고 진단하였다. 국립중앙도서관의 한 보고서에서도 데이터의 풍부한 연결망이 예상치 못한 발견을 촉진하는 환경을 조성한다고 서술한 바 있다(박진호 외, 2025). 즉, 세렌디피티는 개체와 관계의 풍부함을 조건으로 볼 수 있다. 다만 이러한 서술이 발행된 데이터에서도 성립하는지는 확인되어야 한다. 우연한 발견이 개체와 관계의 풍부함을 조건으로 한다면, 개체와 관계를 데이터 조직의 중심에 놓는 검색이 무엇인지를 밝힐 필요가 있다.

시맨틱 검색은 의미와 맥락을 가지고 하는 검색으로 짧게 소개되지만, 무엇이 시맨틱 검색인지에 대한 정확한 정의는 아직 없다(Bast et al., 2016). Guha et al.(2003)은 검색 질의가 가리키는 대상에 근거하여 여러 출처에서 모은 관련 데이터로 전통적인 검색 결과를 보강하는 시스템을 시맨틱 검색으로 서술하였고, Balog(2018)는 개체와 그 속성, 관계를 중심으로 정보를 조직하고 접근하는 검색 패러다임을 개체 중심 검색으로 정의하였다. 이를 바탕으로 본 연구는 시맨틱 검색을 개체와 그 관계로 조직된 데이터 위에서 수행되는 검색으로 정의한다. 개체는 물리적으로 존재하는 사물이나 추상적으로 존재하는 개념으로, URI를 가지며 고유하게 식별될 수 있는 것을 가리킨다.

본 연구는 시맨틱 검색을 탐색과 추론 작업으로 나누어 조작적으로 정의한다. 이 구분은 SPARQL 질의 언어가 결과를 반환하는 2가지 방식에 대응한다. 즉, 명시된 트리플만 맞춰 보는 방식과 선언으로부터 도출된 트리플까지 쓰는 방식이다(W3C, 2013a; 2013b). 본 연구는 질의 언어로 관계를 따라가고 도출하는 범위까지를 확인하며, 이용자가 화면에서 경험하는 검색 기능은 다루지 않는다.

탐색은 관계를 따라 옮겨 가며 연결된 자원을 확인하는 것이다. 소설 한 편에서 그 저자로 옮겨 가고, 저자가 쓴 다른 소설로, 다시 그 소설의 주제로 이어 가는 식이다. 질의 언어에서 탐색은 트리플 패턴을 맞춰 보는 그래프 패턴과 관계를 여러 단계 잇는 속성 경로로 조작된다. 그래프 패턴은 주체와 속성, 객체로 이루어진 트리플 패턴을 하나 이상 조합하여 그래프 안에서 일치하는 서브그래프를 찾는 방식이고, 속성 경로는 속성이 놓이는 자리에 관계를 여러 단계 이어 가거나 방향을 바꾸어 따라가는 경로를 적는다(W3C, 2013a; 2013b). 이렇게 적혀 있는 트리플을 맞춰 보는 것을 부분 그래프 대조 또는 단순 엔테일먼트(simple entailment)라 한다.

추론은 온톨로지를 통해 명시적으로 표현되지 않은 관계를 도출하는 것이다. 사람 사이에 부모 관계가 있고, 부모의 부모는 조부모라는 규칙이 선언되어 있으면, 적혀 있지 않은 조부모 관계를 이끌어 낼 수 있는 것과 같다. 질의 언어에서 추론은 엔테일먼트 레짐(entailment regime)의 적용에 해당한다. 엔테일먼트 레짐은 질의 평가에서 어떤 도출 관계를 사용할 수 있는지를 정한 규정으로(W3C, 2013a), 명시된 트리플을 맞춰 보는 것이 아니라 그 규칙(regime)에 따라 논리적으로 참이 되는 것을 추론(entailment)하여 질의 결과에 포함한다. 탐색이 명시된 관계를 따라간 것이라면 추론은 선언에 따라 도출된 것이라는 점에서 탐색과 추론은 구분된다.

3. 링크드 데이터 발행과 활용의 요건

시맨틱 검색의 제공은 링크드 데이터가 어떻게 발행되어 있는가에 달려 있다. Berners-Lee(2006)는 링크드 데이터의 발행 원칙 4가지를 제시하였다. URI로 이름을 붙이고, HTTP URI를 써서 그 이름을 조회할 수 있게 하며, 조회했을 때 RDF와 SPARQL 같은 표준에 따라 유용한 정보를 제공하고, 더 많은 것을 발견할 수 있도록 다른 URI로의 링크를 포함하라는 것이다. 넷째 원칙이 말하는 링크의 상대에는 다른 기관이 발행한 데이터도 포함된다.

4가지 원칙은 링크드 데이터를 어떻게 발행할 것인가를 정한 것이고, 본 연구에 필요한 것은 국가서지 LOD의 데이터가 시맨틱 검색을 지원하는지 판단할 기준이다. 그래서 원칙에 따라 발행과 활용의 요건 4가지를 정리하여 판단에 활용하였다.

첫째 요건인 개체의 식별은 조회할 수 있는 URI로 이름을 붙이라는 첫째와 둘째 원칙에 대응한다. 도서, 저자, 주제와 같은 주요 자원이 문자열이 아니라 URI를 가진 개체로 제공되는지 확인한다. URI로 식별되는 자원만이 다른 트리플의 주체나 객체로 다시 등장하여 관계를 따라가는 경로의 노드로 기능한다. 문자열 값으로만 제공되는 요소는 결과 표시나 키워드 검색에 쓰일 수 있으나 관계 확장의 출발점이 되지 못한다. 여기에 본 연구는 자원에 클래스(rdf:type)가 부여되어 있는지를 더한다. 클래스가 없으면 그 개체를 유형으로 묶어 찾을 수 없기 때문이다.

둘째 요건인 의미 모델의 정의는 정보 제공의 표준으로 RDF를 명명한 셋째 원칙에 대응하며, 개체와 관계가 어떤 모델로 기술되는지 확인한다. RDFS는 클래스와 관계의 계층을 규정하고, 어떤 유형의 주체와 객체에 그 관계를 쓸 수 있는지를 정의역과 치역으로 밝힌다. 저자 관계(dcterms:creator)라면 정의역은 도서, 치역은 저자가 된다. OWL은 관계 사이의 대응을 더한다. 역관계나 대칭 선언이 있으면 데이터에 적히지 않은 방향의 관계도 선언에서 도출된다. 점검은 표준 어휘의 도입 여부에서 그치지 않고, 선언한 어휘와 클래스가 실제 데이터에서 쓰이는 것과 맞는지까지 확인한다.

셋째 요건인 관계를 통한 연결은 다른 URI로의 링크를 포함하라는 넷째 원칙에 대응하며, 개체와 개체가 URI로 이어지는지 확인한다. 이때 상대는 같은 데이터셋 안의 개체일 수도 있고 다른 기관이 발행한 데이터의 개체일 수도 있다. 넷째 원칙이 안팎을 가르지 않기 때문이다. 점검 항목은 3가지이다. 관계가 URI로 성립하는지와 그 값이 문자열로 저장되어 있지는 않은지는 넷째 원칙에 따른 항목이고, 어느 방향에서든 따라갈 수 있는지가 경로의 폭을 좌우하므로 정방향에 대응하는 역방향이 갖추어져 있는지를 본 연구에서 추가하였다.

넷째인 활용의 요건은 앞에서 정의한 시맨틱 검색, 즉 탐색과 추론의 수행이다. 앞의 3가지 요건이 데이터가 개체와 관계로 조직되어 있는지 확인한다면, 이 요건은 그 위에서 관계를 따라 개체를 찾는 탐색과 추론이 실제로 수행될 수 있는지를 확인한다. 질의가 성립하더라도 관계의 종류가 단순하면 경로도 그만큼 좁아지므로, 관계의 종류와 변별력을 수행 범위의 조건으로 확인

한다. 4가지 요건 가운데 충족되지 못한 항목은 개선의 검토 대상이 된다.

Ⅲ. 국가서지 LOD의 데이터 구조 분석

1. 분석 대상 사례 데이터 확보

본 연구의 분석 대상은 국가서지 LOD가 다운로드 형식으로 제공하는 데이터셋이다. 따라서 여기서 기술하는 구조와 그에 대한 판정은 이 데이터셋의 사례에 한정된다.

국가서지 LOD 홈페이지는 다운로드 데이터셋을 서지데이터(오프라인 자료와 온라인 자료), 자료유형별 서지데이터(일반도서와 학위논문 등 5종), 전거데이터(개인명, 단체명, 주제명), 기타 데이터로 나누어 제공한다. 본 연구는 이 가운데 일반도서 데이터셋과 개인명, 단체명, 주제명 전거데이터셋의 4종을 범위로 삼아, 2026년 4월 1일 기준으로 생성된 데이터셋 파일을 2026년 6월 14일에 다운로드하여 데이터를 확보하였다. 본문의 수치는 이 판본 기준이며, 이후 생성된 판본이나 검색 서비스에서 조회되는 값과는 다를 수 있다.

본 연구는 한국 현대 소설(KDC 분류기호 기준 813.7)로 분류된 도서를 분석 사례로 삼았다. 일반도서 데이터셋의 레코드 3,884,914건에는 오프라인 자료의 서지레코드와 전자책 등 온라인 자료의 레코드가 함께 들어 있다. 본 연구는 이 가운데 오프라인 자료의 서지레코드로서 분류기호 속성(국립중앙도서관 자체 온톨로지의 nlon:kdc) 값에 813.7이 포함되는 도서를 확인하고, 해당하는 개체를 고유 URI 단위로 병합하는 방식으로 사례를 확정하였다. 이 과정을 거쳐 확정된 사례는 도서 53,548건이다.

사례 데이터의 수집은 도서에서 출발하였다. 추출한 도서의 RDF 데이터에서 저자와 주제로 향하는 URI 연결이 확인되었다. 이에 따라, 저자와 주제 개체도 분석 대상에 포함하기 위하여 개인명과 단체명, 주제명 전거데이터셋을 함께 내려받아 도서 개체와 묶었다. 어휘와 클래스의 선언을 사례의 실제 사용과 대조하기 위하여 국가서지 온톨로지 문서도 함께 확보하였다. 정제는 개체를 잃지 않는 범위에서 하였다. 분류 표기가 표준형이 아닌 개체는 자동 보정 없이 표기 그대로 추출하였고, 같은 개체가 여러 파일에 나타나는 경우에는 원자료의 관계와 값을 유지한 채 URI를 기준으로 통합하였다. 도서가 저자로 가리키는 URI는 13,203개이나 그 가운데 5개는 전거 데이터셋에 대응하는 개체가 없어 서브그래프가 수집되지 않았다. 이후의 저자 관련 값은 서브그래프가 확보된 개체를 기준으로 한다. 같은 이름을 가진 저자가 서로 다른 URI로 존재하는 경우가 있으나, 저자 URI를 그대로 분석 단위로 삼으며 동일 인물 여부는 판별하지 않는다.

국가서지 LOD는 SPARQL Endpoint를 제공하여 값을 조회할 수 있다. 그러나 추론을 확인하려면 엔테일먼트 레짐을 적용해야 하는데, 국가서지 LOD의 SPARQL Endpoint에서는 어떤 레짐이 적용되는지

알 수 없고 연구자가 임의로 설정할 수도 없다. 그래서 연구자가 직접 설정할 수 있는 저장소에 적재하여 질의를 수행하고자 RDF와 SPARQL을 지원하는 그래프 데이터베이스 프로그램인 Ontotext GraphDB 11.4.1을 사용하였다. RDF 그래프는 같은 내용을 서로 다른 문법으로 기술하는 여러 직렬화 형식으로 저장할 수 있으며, 국가서지 LOD도 데이터셋을 TTL, RDF/XML, JSON-LD의 3가지 형식으로 제공한다. 다운로드한 파일을 GraphDB에 형식별 저장소로 적재하고 같은 질의로 값을 대조한 뒤, TTL 저장소를 분석의 기준으로 삼았다. 본문의 값은 별도의 표시가 없는 한 TTL 기준이다.

2. 분석 방법과 SPARQL 질의 구성

분석의 기준은 Ⅱ장에서 링크드 데이터의 발행 원칙을 바탕으로 정리한 4가지 요건이다. 첫째 요건인 개체의 식별은 분석 대상인 도서와 저자, 주제가 각각 URI를 가진 개체로 세워져 있는지와 그 개체에 클래스(rdf:type)가 부여되어 있는지를 확인하였다. 둘째 요건인 의미 모델의 정의는 온톨로지 문서가 선언한 속성과 클래스를 사례에 실제로 나타나는 것과 대조하여 확인하였다. 셋째 요건인 관계를 통한 연결은 주체와 객체의 연결에서 객체가 URI를 가지는 자원인지 리터럴 값으로 표시되는지를 구분하고, 정방향 관계에 대응하는 역방향이 채워져 있는지를 확인하였다. 넷째 요건인 시맨틱 검색의 수행은 특정한 질의 형식에서 반환되는 결과의 특성을 확인하였다. 앞의 3가지 요건은 발행된 데이터가 갖춘 상태를 파악하는 것으로 연구문제 1에 대응하여 본 장에서 다루고, 넷째 요건과 요건 점검에서 드러난 미충족에 대한 개선은 다음 장에서 다룬다. 본 장에서 다루는 3가지 요건은 차례로 3절 그래프 기본 구조, 4절 온톨로지와 클래스 구성, 5절 관계 구성과 외부 연결에서 확인한다.

본 장에서는 사례의 기본 구조를 확인하기 위한 질의와 연결과 값의 결손을 확인하기 위한 질의를 구성하였다. 분석에 사용한 데이터와 질의, 스크립트, 그 결과는 공개 저장소[1]에 게시하였고, 본문의 값은 보존된 질의를 다시 실행하여 대조하였다.

각 기술이 맡는 역할은 다음과 같다. 집계는 기본 그래프 패턴과 그 질의 결과의 수를 세는 것으로 개체가 URI로 세워져 있는지와 개체 사이의 관계가 URI로 이어지는지를 확인하는 데 사용하였다. 속성 경로는 관계를 여러 단계 따라가는 확장이 어디까지 이르는지를 확인하는 데 사용하였다. 집계는 별도의 언급이 없는 한 관계의 상대편이 사례 밖 개체인 경우까지 포함하며, 사례 안에서 클래스가 확인되는 개체 사이만 세는 실질 연결과 구분한다.

3. 그래프 기본 구조

그래프의 기본 구조는 도서 한 건의 개체에서 출발하여 확인한다. 사례에 포함된 도서 「바깥은

1) https://github.com/pnumksong/nlk-lod-semantic-search

여름」(김애란 지음, 문학동네, 2017)을 주체로 하는 트리플은 43건이고 그것이 이 도서의 서브그래프를 이루며, 이 43건에 나타나는 속성은 33종이다. 객체 자리를 보면 URI 참조가 12건, 리터럴 값이 31건이다. URI 참조 12건은 번역본 연결 5건과 타입 값 3건, 그리고 저자 연결, 주제 연결, 번역 원작 연결, 자료 유형 연결 각 1건이다. 리터럴 값 31건은 표제와 발행사항, 분류기호, 소장사항 등과 같은 속성의 값이 문자열이라는 것이다.

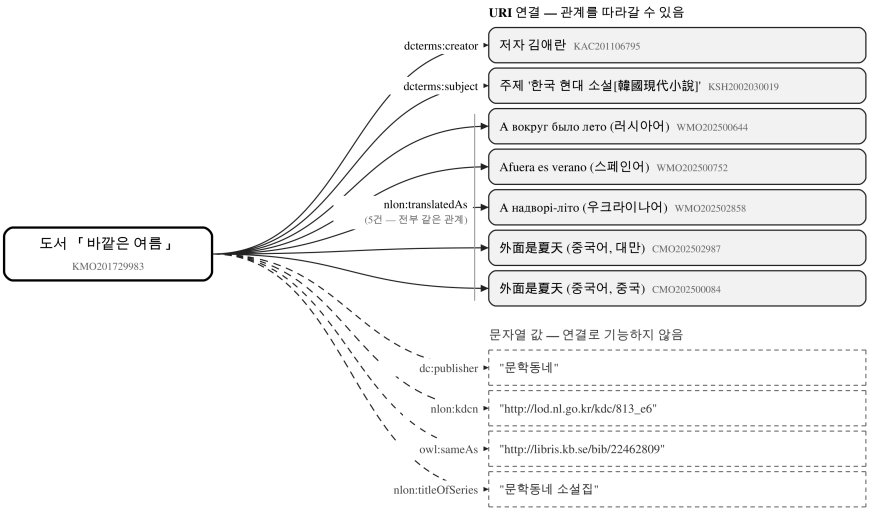
관계를 따라갈 수 있는 URI 속성과 그렇지 못한 리터럴 속성이 함께 있는 이 구성은 도서 전체에서도 확인된다. 각 도서 개체를 서술하는 속성을 정리하면 〈표 1〉과 같다. nlon:Book 타입이 부여된 53,547건을 기준으로, 리터럴 속성은 8,000건 이상인 32종을, 관계 속성은 입력 건수와 무관하게 15종 전량을 입력 건수 순으로 표시하였으며, 각 속성이 연결로서 기능할 수 있는지를 값 형식을 통해 파악하였다. 속성의 설명과 원천 필드는 국가서지 LOD 홈페이지가 제공하는 RDF 매핑을 따랐으며(국립중앙도서관, 발행년불명), 타입 값(rdf:type)은 클래스 구성에서 다루므로 제외하였다. 도서에는 표제 같은 문자열 값과 저자, 주제, 번역본으로의 URI 연결이 함께 있으며, 외부 식별자와 출판사, 총서명처럼 연결로 쓰일 수 있는 값들이 문자열로 저장되어 있어 연결로 기능하지 못한다. 이처럼 문자열로만 입력된 값은 키워드 검색에는 활용될 수 있으나, 동명이인이나 이형표기 등 개체 식별이 불가능하여 개체 사이의 관계를 통해 저자나 주제, 관련 도서로 확장하는 탐색의 출발점이 되지 못한다.

〈표 1〉 도서 개체의 속성 구성

속성	설명	값 형식	건수	비고
nlon:localHolding	소장사항	리터럴	176,632	KORMARC 049 필드
bf:extent	형태사항	리터럴	133,430	KORMARC 300 필드
bibo:isbn	국제표준도서번호	리터럴	102,309	KORMARC 020 필드
dc:creator	표목의 저자 이름	리터럴	87,899	KORMARC 100, 110, 700 필드
dcterms:alternative	여러 형태의 표제	리터럴	71,358	KORMARC 242, 246 필드
dcterms:subject	주제 개체 연결	URI	66,859	KORMARC 600 필드 등
dc:publisher	발행처	리터럴	60,306	KORMARC 260 필드
nlon:kdc	한국십진분류기호	리터럴	54,167	KORMARC 056 필드
nlon:itemNumberOfNLK	청구기호의 도서기호	리터럴	53,854	KORMARC 052 필드
nlon:classificationNumberOfNLK	청구기호의 분류기호	리터럴	53,795	KORMARC 052 필드
nlon:publicationPlace	발행지	리터럴	53,739	KORMARC 260 필드
dcterms:title	본표제	리터럴	53,650	KORMARC 245 필드
dcterms:issued	발행년	리터럴	53,553	KORMARC 260 필드
rdfs:label	본표제	리터럴	53,547	KORMARC 245 필드
bf:language	언어 개체 주소	리터럴(URI 형식)	53,547	-
bf:place	발행국 개체 주소	리터럴(URI 형식)	53,547	-
nlon:datePublished	데이터 생성 일시	리터럴	53,547	시스템 값
nlon:typeOfData	자료 유형 개체 연결	URI	53,547	-
nlon:issuedYear	발행 연도	리터럴	53,490	-
nlon:editionOfKDC	KDC 판표시	리터럴	53,485	KORMARC 056 필드
nlon:kdcn	KDC 분류 개체 주소	리터럴(URI 형식)	53,336	-
dcterms:creator	저자 개체 연결	URI	50,295	KORMARC 100, 110, 700 필드
rdfs:seeAlso	외부 서지와 백과의 웹 주소	리터럴	39,528	-
nlon:titleOfSeries	총서표제	리터럴	32,331	KORMARC 440 필드
nlon:remainderOfTitle	표제관련정보	리터럴	32,117	KORMARC 245, 247 필드
nlon:ddc	듀이십진분류기호	리터럴	28,353	KORMARC 082 필드
nlon:editionOfDDC	DDC 판표시	리터럴	28,131	KORMARC 082 필드

속성	설명	값 형식	건수	비고
dcterms:tableOfContents	내용주기	리터럴	21,851	KORMARC 505 필드
nlon:uniformTitleOfSeries	총서의 통일표제	리터럴	21,353	KORMARC 830 필드
nlon:volumeOfSeries	총서번호	리터럴	14,909	KORMARC 440, 490, 830 필드
dcterms:description	일반주기	리터럴	14,480	KORMARC 500 필드 등
nlon:cip	출판시도서목록(CIP) 제어번호	리터럴	13,832	KORMARC 023 필드
nlon:volumeOfNLK	청구기호의 권, 연차기호	리터럴	10,589	KORMARC 052 필드
nlon:ddcn	DDC 분류 개체 주소	리터럴(URI 형식)	8,656	–
owl:sameAs	외부 서지 식별자	리터럴(URI 형식)	8,284	–
nlon:translatedAs	번역본 개체 연결	URI	937	KORMARC 767 필드
bibo:translationOf	원저 개체 연결	URI	763	KORMARC 765 필드
dcterms:hasVersion	이판 개체 연결	URI	256	KORMARC 775 필드
nlon:preceding	선행저록 개체 연결	URI	127	KORMARC 780 필드
nlon:succeeding	후속저록 개체 연결	URI	53	KORMARC 785 필드
dcterms:hasFormat	기타형태 개체 연결	URI	30	KORMARC 776 필드
nlon:series	총서 개체 연결	URI	24	–
dcterms:relation	관련 자료 개체 연결	URI	23	KORMARC 787 필드
dcterms:isPartOf	상위 자료 개체 연결	URI	19	KORMARC 490, 773 필드
nlon:hasSupplement	보유판 개체 연결	URI	10	KORMARC 770 필드
nlon:supplementTo	모체 자료 개체 연결	URI	8	KORMARC 772 필드
dcterms:hasPart	구성단위 개체 연결	URI	1	KORMARC 490, 774 필드

〈그림 1〉은 위의 개체에서 주요 속성을 발췌하여, 이 도서가 무엇과 URI로 연결되어 있고 무엇이 문자열로 남아 있는지를 보인 것이다. 이 도서는 저자 개체(김애란)와 주제 개체('한국 현대 소설'), 그리고 여러 언어의 번역본 개체 5건과 URI로 연결되어 있어, 그 연결을 따라 같은 저자의 다른 도서나 같은 주제의 도서, 이 책의 번역본으로 옮겨 갈 수 있다. 반면 출판사는 '문학동네'라는 문자열로만 있고, KDC 분류 개체를 가리키는 값과 스웨덴 국립도서관 Libris의 같은 책을 가리키는 owl:sameAs 값은 URI 형식을 갖추고도 문자열로 저장되어 있으며, 총서명 '문학동네 소설집'도 문자열로만 적혀 있다. 그림에는 개체의 속성 33종 가운데 URI 연결 3종과 문자열 값 4종, 합하여 7종을 발췌하였다.



〈그림 1〉 도서「바깥은 여름」개체의 URI 연결과 문자열 값

도서에서 URI로 이어지는 상대는 번역본과 연관저록 같은 다른 도서, 저자, 주제, 자료 유형이다. 이 가운데 자료 유형은 사례를 일반도서로 한정한 까닭에 도서 53,547건 가운데 53,545건이 같은 값을 가리켜 도서를 서로 가르지 못한다. 언어와 발행국, 분류기호도 URI 형식을 갖추고 있으나 값이 문자열로 저장되어 연결로 기능하지 않는다. 관계 속성 가운데 주제, 자료 유형, 저자를 제외한 12종은 1,000건 미만으로 일부만 입력되어 있다. 도서 밖의 다른 개체로 나가는 연결은 저자와 주제 2가지다. 현행 국가서지 LOD 구조에서 도서 개체의 저자와 주제, 번역본 등 연관저록으로 관계를 확장하여 탐색할 수 있지만, 그 외에 같은 출판사나 같은 총서의 다른 도서 등으로 관계를 확장하여 탐색하는 것은 제한적이다.

도서에서 저자로의 연결에는 dcterms:creator 속성을 사용하며, 이 연결이 가리키는 저자 개체는 13,198개로 식별된다. 그 속성 구성을 정리하면 〈표 2〉와 같다. 저자에는 이름과 도서 연결에 더해 직업과 활동 분야 같은 값이 있다. 표에는 서브그래프가 확보된 저자 13,198개를 기준으로, 타입 값을 제외하고 5,000건 이상인 속성 16종을 실었다.

〈표 2〉 저자 개체의 속성 구성

속성	설명	값 형식	건수	비고
nlon:create	저작 연결	URI	450,053	KORMARC 전거 100, 110 필드
owl:sameAs	외부 전거 식별자	리터럴(URI 형식)	29,516	KORMARC 전거 024 필드
dc:source	정보원	리터럴	26,397	KORMARC 전거 670 필드
schema:jobTitle	직업	리터럴	26,255	KORMARC 전거 374 필드
nlon:fieldOfActivity	분야	리터럴	23,197	KORMARC 전거 372 필드
nlon:associatedLanguage	관련언어	리터럴	13,828	KORMARC 전거 377 필드
skos:altLabel	이형 이름 표기	리터럴	13,405	-
rdfs:label	이름	리터럴	13,198	KORMARC 전거 100, 110 필드
dcterms:modified	최종수정일	리터럴	13,198	KORMARC 전거 005 필드
skos:prefLabel	우선 이름 표기	리터럴	13,198	-
foaf:name	개인명, 단체명	리터럴	13,198	KORMARC 전거 100, 110 필드
nlon:isni	기타 표준부호(ISNI)	리터럴	13,082	KORMARC 전거 024 필드
nlon:corporateName	관련 단체명	리터럴	10,801	KORMARC 전거 510 필드
schema:gender	성별	리터럴	7,615	KORMARC 전거 375 필드
nlon:birthYear	출생년	리터럴	5,845	KORMARC 전거 100 필드
schema:birthPlace	관련장소	리터럴	5,705	KORMARC 전거 370 필드

저자 개체에서 주로 사용된 속성 16종 중에서 관계 확장을 통한 탐색으로 기능할 수 있는 것은 도서 연결뿐이다. 도서에서 저자는 2가지 방법으로 표시된다. 더블린 코어(Dublin Core) 기본 요소의 이름 문자열(dc:creator)과 확장 어휘의 URI 연결(dcterms:creator)이다. 예컨대 도서 「바깥은 여름」은 dc:creator에 '김애란'이라는 문자열을 가지며, dcterms:creator로는 같은 이름의 저자 개체를 URI로 가리킨다. 저자 이름이 문자열 값으로만 있고 저자 URI와의 연결이 없는 도서가 17,412건으로, 저자명 문자열을 가진 도서 52,840건의 32.95%에 해당한다. 이 도서들에서 저자는 결과 표시에는 쓰일 수 있으나 관계 확장의 출발점이 되지 못한다. 반면 저자 URI와 이름 문자열이 함께 적힌 도서는 35,428건이며, 이 가운데 99.73%(35,331건)에서 도서에 적힌 이름과

연결된 저자 개체의 이름(foaf:name)이 일치한다. 다만 이것이 문자열만 있는 도서를 저자 개체로 이을 수 있다는 뜻은 아니다. 같은 이름의 다른 저자, 즉 동명이인 저자가 있어, 이름만으로는 어떤 저자 개체를 가리키는지 정해지지 않기 때문이다.

도서에서 주제로의 연결에는 dcterms:subject 속성을 사용하며, 이 연결이 가리키는 주제 개체는 815개로 식별된다. 주제 개체를 서술하는 속성을 정리하면 〈표 3〉과 같다. 주제에는 도서로의 연결과 주제 사이의 상위, 하위, 관련 연결이 있으며, 주제 사이 연결의 구성은 5절에서 다룬다.

〈표 3〉 주제 개체의 속성 구성

속성	설명	값 형식	건수	비고
nlon:isSubjectOf	도서 연결	URI	1,443,908	KORMARC 600 필드 등
skos:related	관련어	URI	7,512	시소러스 관계 RT
skos:narrower	하위개념어	URI	5,417	시소러스 관계 NT
skos:altLabel	유사어, 동의어와 외국어 표기	리터럴	3,493	시소러스 관계 UF 등
owl:sameAs	주제명 기반 URI와의 동일성	URI 815, 리터럴 55	870	–
rdfs:label	주제명	리터럴	815	–
skos:inScheme	개념 체계 소속	URI	815	–
skos:broader	상위개념어	URI	748	시소러스 관계 BT
skos:prefLabel	우선어	리터럴	565	시소러스 관계 USE
skos:definition	정의문	리터럴	213	–
skos:closeMatch	유사 일치	리터럴	117	–
rdfs:seeAlso	외국어 용어 연계	URI	44	시소러스 언어 대응

〈표 1〉~〈표 3〉의 건수는 해당 클래스의 개체가 가진 트리플 수이며, 값 형식이 URI와 리터럴로 갈리는 속성은 URI 객체를 가진 트리플만 셈하였다. 도서의 속성은 nlon:Book 타입이 부여된 53,547건을 기준으로 집계한 것이어서, 그 타입이 없는 1건이 빠져 전체 도서 53,548건 기준의 값과 그 도서가 가진 값의 수만큼 차이가 난다.

사례 그래프 전체의 규모는 다음과 같다. 적재된 트리플은 4,133,636건이고, 주체 자리에 나타나는 고유 개체는 67,561개로 도서 53,548건과 저자 13,198개, 주제 815개를 합한 수와 같다. 속성은 120종이며, 객체 자리는 URI 참조 2,256,638건과 리터럴 값 1,876,998건으로 나뉜다. 도서와 저자, 주제 사이의 관계는 각각 속성 하나뿐이어서 세분화되어 있지 않다. 저자 연결의 3분의 1은 문자열로만 남아 있으며, 출판사와 총서명처럼 연결로 쓰일 수 있는 값도 문자열로 저장되어 관계 확장의 출발점이 되지 못한다.

4. 온톨로지와 클래스 구성

사례의 속성 120종은 네임스페이스 11종에 걸쳐 있다. 온톨로지 문서의 선언과 사례의 사용을 대조하면 〈표 4〉와 같다. 이 가운데 자체 온톨로지인 nlon의 속성이 70종으로 가장 많으며, 전체 트리플의 68.34%를 차지한다.

〈표 4〉 네임스페이스별 선언과 속성 구성

네임스페이스	온톨로지 선언	속성(종)	트리플(건)	URI 객체(건)	리터럴 객체(건)
nlon	○	70	2,824,790	1,948,775	876,015
dcterms	○	16	347,879	117,483	230,396
BIBFRAME(id.loc.gov)	X	3	240,528	0	240,528
rdf	○	1	187,464	174,266	13,198
dc	○	5	174,920	0	174,920
bibo	○	6	111,961	763	111,198
rdfs	○	3	107,590	44	107,546
skos	○	8	45,483	14,492	30,991
schema.org	X	5	41,088	0	41,088
owl	○	1	38,671	815	37,856
foaf	○	2	13,262	0	13,262
voaf	○	-	-	-	-
vann	○	-	-	-	-
cc	○	-	-	-	-
vs	○	-	-	-	-
nlk	○	-	-	-	-
xsd	○	-	-	-	-
geo	○	-	-	-	-
org	○	-	-	-	-
BIBFRAME(bibframe.org)	○	-	-	-	-
합계	18종(미선언 2종)	120	4,133,636	2,256,638	1,876,998

온톨로지 문서가 선언한 네임스페이스는 문서 기술용을 포함하여 18종이지만, 사례 데이터에 속성으로 나타나는 것은 11종이고, 둘이 겹치는 것은 9종이다. 선언만 있고 사례에서 쓰이지 않는 9종이 모두 문제인 것은 아니다. 어휘 문서를 기술하는 4종(voaf, vann, cc, vs)과 식별 공간 2종(nlk, xsd)은 성격상 데이터의 속성 자리에 나타나지 않는다. 선언과 사용의 어긋남은 나머지 3종에서 나타난다. 지리 좌표(geo)와 조직(org)의 어휘 2종은 선언되고도 데이터에 나타나지 않는다. 도서관 서지정보를 링크드 데이터로 표현하기 위하여 미국의회도서관이 주도하여 개발한 어휘인 BIBFRAME은 문서가 선언한 네임스페이스(bibframe.org)와 데이터가 실제로 쓰는 네임스페이스(id.loc.gov)가 달라, 선언만으로는 이 어휘가 데이터에서 쓰이는지 확인할 수 없다. 반대로 사례가 쓰는 schema.org 어휘는 온톨로지 문서에 선언되어 있지 않다.

〈표 5〉 사례에 나타나는 클래스

계열	클래스	온톨로지 선언	값 형식	개체 수
도서	nlon:OfflineMaterial	○	URI	53,548
	nlon:Book	○	URI	53,547
	bibo:Book	X	URI	53,547
저자	nlon:Author	○	URI	12,807
	foaf:Person	X	리터럴	13,040
	foaf:Organization	X	리터럴	158
주제	nlon:Concept	○	URI	815
기타자료	nlon:NonBook	○	URI	1
	nlon:Score	○	URI	1

클래스 구성에서도 선언과 사용이 일치하지 않는다. 사례에 나타나는 클래스는 〈표 5〉의 9종이며, 그 가운데 6종이 온톨로지 문서에서 선언된 클래스이고 bibo:Book과 foaf:Person, foaf:Organization 3종은 온톨로지에 선언되지 않은 것이다. 온톨로지 문서가 선언한 클래스는 자료 유형 계열과 분류, 국가, 장소, 도서관 등 31종인데, 사례 데이터에 타입 값으로 나타나는 것은 그 6종뿐이고 나머지 25종은 나타나지 않는다. 나타나지 않는 25종 가운데 국가와 장소, 자료 유형 코드 5종은 2026년 4월 1일 기준 데이터셋 4종 전체에서도 나타나지 않는다. 한 도서에 nlon:OfflineMaterial과 nlon:Book, bibo:Book이 함께 부여되어 있어 표의 개체 수를 더하면 실제 개체 수를 넘으며, foaf:Person과 foaf:Organization은 타입 값이 문자열로 저장되어 있다.

선언된 클래스와 사례에 나타나는 클래스를 견주면 대비가 드러난다. 선언된 클래스 31종은 계열마다 여럿인데 사례에 나타나는 것은 계열별로 한두 종으로, 사례의 개체를 실질적으로 조직하는 클래스 구분은 도서, 저자, 주제의 3가지뿐이다. 자료 유형 계열의 대부분이 나타나지 않는 것은 일반도서로 사례를 한정한 데서 비롯되지만, 분류와 총서, 국가, 장소처럼 일반도서에도 부착될 수 있는 클래스는 선언만 있고 사례에서 개체로 나타나지 않는다.

타입을 부여하는 방식에도 결함이 있다. 저자 URI 13,198개는 자원의 유형을 적는 타입 값(rdf:type)이 문자열로 저장되어 있다. 클래스 이름인 foaf:Person이 URI가 아니라 문자열로 적혀 있는 것이다. 문자열은 클래스로 기능하지 않으므로, foaf:Person에 속하는 자원을 찾는 질의에 이 저자들은 나타나지 않는다. 이 가운데 12,807개는 nlon:Author라는 URI 클래스를 함께 가지지만, 나머지 391개는 URI 클래스가 부여되어 있지 않다. 타입 값이 문자열인 것은 원본 데이터셋에서도 같아, 추출이나 통합 과정에서 생긴 것이 아니라 원천 데이터셋의 특성이다. 타입 부여의 누락은 도서에서도 나타난다. 도서 53,548건 가운데 1건은 nlon:Book과 bibo:Book 어느 쪽으로도 타입이 부여되어 있지 않다.

dc, schema.org, foaf, BIBFRAME(id.loc.gov) 어휘는 객체가 전량 리터럴이다(〈표 4〉 참조). 다만 리터럴 값 자체가 결함은 아니다. 표제나 이름처럼 값을 문자열로 담는 것은 리터럴 속성의 본래 용법이며, 문제가 되는 것은 URI로 식별되는 자원을 가리켜야 할 자리가 문자열로 남는 경우다. 같은 이름을 가진 자료가 여럿일 수 있어, 이름 문자열만으로는 가리키는 대상이 식별되지 않기 때문이다. 온톨로지의 선언과 사례 데이터가 어긋나는 방식은 2가지이다. 분류는 Classification 클래스와 분류 연결 관계가 온톨로지에 선언되어 있음에도 사례에서 개체가 아니라 문자열 값으로만 부착된다(분류 계열 속성 6종 226,134건 전량 리터럴). 총서도 같은 형태로, 총서 클래스(nlon:Collection)와 총서 연결 관계가 선언되어 있으나 사례에는 그 클래스의 개체가 하나도 없다. 출판사는 대응하는 클래스 선언 자체가 없이 문자열 값으로만 존재한다(60,307건, 고유 문자열 6,123종). 즉, 분류와 총서는 선언된 구조를 데이터가 채우지 못한 것이고 출판사는 선언 자체가 없는 것으로, 값의 저장 형식이 문제인 경우와는 구별된다.

5. 관계 구성과 외부 연결

본 절은 개체와 개체가 어떤 관계로 이어지는지 확인한다. 사례 안에서 URI 클래스가 확인되는 개체 사이의 연결을 실질 연결로 셈하며, 관계의 상대편이 사례 밖 개체인 경우까지 포함하는 전량 집계와 구분한다. 속성 120종 가운데 개체를 URI로 잇는 것은 27종이고, 자원의 유형을 기술하는 rdf:type 1종을 제외한 나머지 92종은 리터럴만을 객체로 가진다. 실질 연결의 관계별 규모는 〈표 6〉과 같다. 도서와 주제를 잇는 dcterms:subject, 도서와 저자를 잇는 dcterms:creator가 대부분을 차지하고, 도서와 도서, 주제와 주제를 연결하는 관계는 규모가 작다. 〈표 6〉의 저자 축에서 도서와 저자를 잇는 관계는 dcterms:creator 하나이다. 직업 값(schema:jobTitle)에 그림, 삽화, 일러스트, 화가 계열의 표기가 있는 저자 URI 2,488개, 번역 계열 376개, 편집 계열 193개가 모두 이 하나의 관계로 연결되어 있어 역할이 다른 저자들이 구분되지 않는다. 도서를 통해 저자와 주제로 이어지는 연결이 성립하는 것은 도서가 저자와 주제를 각각 가리키고 있기 때문이며, 저자와 주제를 직접 잇는 관계는 사례에 없다.

〈표 6〉 개체 사이의 실질 연결 규모

연결	관계	트리플(건)
도서와 주제	dcterms:subject	66,852
도서와 저자	dcterms:creator	49,841
도서와 도서	nlon:translatedAs	544
	bibo:translationOf	532
	dcterms:hasVersion	249
	nlon:preceding	49
	nlon:succeeding	47
주제와 주제	skos:related	470
	skos:broader	243
	skos:narrower	243
합계		119,070

주제 사이의 연결은 지식 조직 체계의 표현을 위한 W3C 표준 어휘인 SKOS(Simple Knowledge Organization System)를 사용한다(W3C, 2009). 주제는 skos:broader, skos:narrower, skos:related 관계를 통해 상위, 하위, 관련 주제로 이어지며, 〈표 6〉의 주제와 주제 축이 그 실질 연결의 규모이다. 이 관계들을 여러 단계 따라가는 확장은 다음 장의 탐색에서 다룬다.

정방향 관계에 대응하는 역방향이 채워진 정도는 〈표 7〉과 같다. 저자나 주제에서 출발하여 도서로 가는 질의는 이 역방향 관계를 따르므로, 도서 쪽에 적힌 연결 가운데 각각 80.46%와 86.54%만을 반환한다. 역방향 결손의 판정 기준은 외부 규범이 아니라 데이터 자체이다. 온톨로지에 역방향의 관계 어휘가 있고 대다수가 실제로 채워져 있으므로, 남은 부분은 설계상의 선택보다는 입력 과정에서 누락된 것에 가깝다. 예를 들어 도서 「축구도 잘해요」는 저자 관계로 김애란을 가리키지만 김애란에서 도서로 향하는 관계에는 이 도서가 없어, 김애란에서 출발한 질의로는 「축구도 잘해요」가 반환되지

않는다. 도서 「왕을 찾아서」는 주제 '한국 현대 소설'을 가리키지만 그 주제에서 도서로 향하는 관계에는 이 도서가 없어, 주제 '한국 현대 소설'에서 출발한 질의로는 「왕을 찾아서」가 반환되지 않는다.

〈표 7〉 정방향 관계와 역방향의 충족

관계	정방향	역방향	정방향 (트리플)	역방향이 함께 있는 트리플	역방향 결손 (트리플)
저자 관계	도서에서 저자로 (dcterms:creator)	저자에서 도서로 (nlon:create)	50,295	40,468 (80.46%)	9,827
주제 관계	도서에서 주제로 (dcterms:subject)	주제에서 도서로 (nlon:isSubjectOf)	66,859	57,859 (86.54%)	9,000

외부 데이터셋으로의 연결은 같은 대상을 가리키는 자원을 잇는 owl:sameAs와 참고할 자원을 잇는 rdfs:seeAlso 속성을 사용한다. 이 2가지 속성을 포함한 주요 속성의 객체 형식은 〈표 8〉과 같다. owl:sameAs가 가리키는 대상 데이터셋은 국제 표준 이름 식별자인 ISNI 13,082건, 국제 전거 파일인 VIAF 11,753건, 국회도서관 4,006건, 스웨덴 Libris 3,571건, 미국 의회도서관 1,360건 등 24종이다.

owl:sameAs 가운데 URI 객체로 저장된 815건은 외부 데이터셋이 아니라 국가서지가 주제마다 함께 부여한 또 하나의 URI를 가리키며, 사례의 주제 815개가 각각 하나씩 갖는다. 주제에는 식별자 기반 URI와 주제명 기반 URI가 함께 부여되어 owl:sameAs로 이어져 있다(주제명 전거데이터셋 레코드 257,698건 전건). 그 상대 URI는 사례 수집 범위 밖이어서 클래스가 확인되는 개체 사이의 연결에는 들어가지 않는다.

〈표 8〉 주요 속성의 객체 형식

속성	URI 객체(건)	리터럴 객체(건)	리터럴 비율	리터럴 값의 내용
owl:sameAs	815	37,856	97.89%	외부 전거와 국가서지의 식별자 URI
rdfs:seeAlso	44	39,980	99.89%	외부 서지와 백과의 URL
bf:extent	0	133,432	100%	분량
bf:language	0	53,548	100%	언어 개체 주소
bf:place	0	53,548	100%	발행국 개체 주소
dc:creator	0	87,899	100%	저자명 문자열

Ⅳ. 시맨틱 검색의 적용 및 개선

1. 시맨틱 검색의 적용 방법

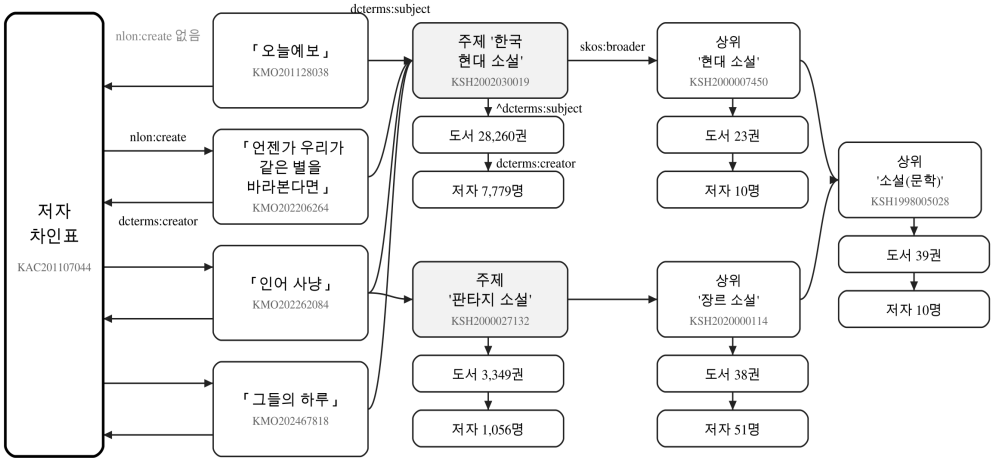
이 장은 연구문제 2와 연구문제 3에 대응하여 앞 장에서 파악한 데이터 구조에서 시맨틱 검색이

어디까지 성립하는지를 확인한다. 앞 장에서 확보한 같은 사례 데이터를 대상으로 앞에서 설명한 SPARQL 질의를 적용하였다. 데이터에 적혀 있는 관계를 질의로 따라가는 탐색은 2절에서, 온톨로지의 선언으로부터 따라 나오는 관계까지 질의의 대상에 넣는 추론은 3절에서 다룬다.

2. 탐색 과정에서의 적용

도서 「인어 사냥」의 저자를 묻는 질의는 차인표를 반환하고, 그 도서의 주제를 묻는 질의는 '한국 현대 소설'과 '판타지 소설'을 반환한다. 저자 차인표에서 출발하여 그의 도서와 주제로 이어지는 연결을 옮기면 〈그림 2〉와 같다. 한 단계의 명시된 관계에서는 연결된 개체와 그 수치를 질의로 확인할 수 있으며, 도서에서 저자로, 도서에서 주제로 따라가는 질의가 모두 결과를 반환한다. 저자와 주제에 모두 URI 연결이 있어 양방향으로 확장할 수 있는 도서는 31,692건으로, 사례 도서 53,548건의 59.18%이다.

관계를 이어 가는 다단계 질의도 성립한다. 도서를 통해 저자와 주제를 함께 묶는 조합이 62,787건이고, 저자에서 그 저자의 도서가 다루는 주제로 이어지는 고유한 쌍이 27,195쌍이다. 여기서 쌍은 경로가 연결하는 주체와 객체의 고유한 짝을 세는 단위다. 예를 들어 도서 「바깥은 여름」과 저자 김애란, 주제 '한국 현대 소설'이 하나의 조합이고, 김애란과 '한국 현대 소설'은 도서를 거쳐 이어지는 하나의 쌍이다. 1단계와 2단계 질의의 결과는 〈표 9〉와 같다. 3단계의 확장도 성립한다. 도서에서 저자를 거쳐 같은 저자의 다른 도서로, 다시 그 도서의 주제로 잇는 질의는 866,294개의 경로(도서, 저자, 다른 도서, 주제의 조합)를 반환한다. 즉, 명시된 관계의 범위 안에서 확장을 제한하는 것은 단계의 수가 아니라 관계의 종류다.



〈그림 2〉 연결 관계 확장의 예시

〈표 9〉 명시된 관계의 질의 집계

구분	질의	묻는 것	값	단위
1단계	도서와 저자의 관계	도서와 저자를 잇는 트리플의 수	50,295	트리플
	도서와 주제의 관계	도서와 주제를 잇는 트리플의 수	66,859	트리플
	저자 URI와 주제 URI를 모두 갖춘 도서	저자로도 주제로도 URI 연결이 있는 도서의 수	31,692	건
2단계	저자와 주제의 고유한 쌍	저자에서 도서를 거쳐 주제로 이어지는 서로 다른 쫙의 수	27,195	쌍
	도서, 저자, 주제의 조합	도서 한 건과 그 저자, 그 주제를 한 묶음으로 셈한 수	62,787	건

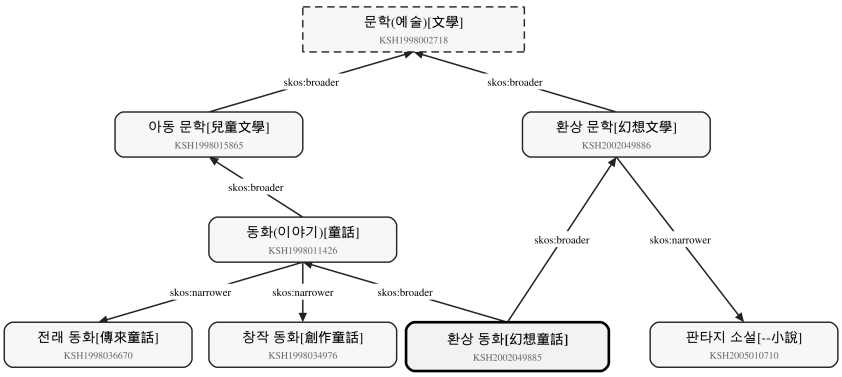
〈표 9〉는 관계 전체를 기준으로 셈하므로, 도서와 저자를 잇는 트리플은 50,295건으로 〈표 6〉의 49,841건보다 454건 많고, 도서와 주제를 잇는 트리플은 66,859건으로 66,852건보다 7건 많다. 이 차이는 URI 클래스가 부여되지 않은 URI로의 연결이며, 주제 축의 7건은 클래스가 부여되지 않은 주제 URI 2개로의 연결이다.

주제 사이의 관계는 속성 경로로 여러 단계 이어 갈 수 있다. 사례와 연결된 주제 815개 가운데 상위 관계를 가진 것은 561개, 하위 관계를 가진 것은 398개, 관련 관계를 가진 것은 707개이다. 관계별 트리플과 경로 확장의 결과는 〈표 10〉과 같다. 경로를 무제한으로 열어도 상위와 하위 방향의 확장은 4단계에서 더 이상 새로운 주제에 도달하지 않는다. 상위 관계와 관련 관계를 함께 열어 경로를 확장하면 89,436쌍이 얻어져, 2가지 관계를 따로 열었을 때의 합보다 2.3배 많다. 즉, 경로를 조합하면 도달 범위가 넓어지나, 주제 그래프 자체의 규모가 작아 경로를 따라 새로 연결되는 주제의 수는 제한적이다. 이 규모는 사례를 하나의 분류 계열로 한정한 결과이고, 데이터 구조의 문제와는 별개이다. 다만 관계의 풍부함을 조건으로 하는 우연한 발견의 관점에서 보면, 이러한 주제 관계의 제한된 규모는 그 조건이 아직 갖추어지지 않았음을 뜻한다.

〈표 10〉 SKOS 주제 관계와 경로 확장

구분	항목	값	단위	관계를 가진 주제 (개)
관계	상위 관계(skos:broader)	748	트리플	561
	하위 관계(skos:narrower)	5,417	트리플	398
	관련 관계(skos:related)	7,512	트리플	707
경로 확장	상위 방향(broader+)	1,109	쌍	–
	하위 방향(narrower+)	6,803	쌍	–
	관련 방향(related+)	37,050	쌍	–
	상위 또는 관련 관계	89,436	쌍	–
심도	최대 심도(상위와 하위 방향)	4	단계	–

주제의 상하위 확장은 이 절에서 속성 경로를 적용하여 수행하였다. 주제 그래프는 상위 주제 아래에 하위 주제가 놓이고 관련 관계가 더해진 구조이다. 예를 들어 '환상 동화'는 '동화'의 하위 주제이고 '동화'는 '아동 문학'의 하위 주제여서, '아동 문학'에서 출발해 2단계를 내려가면 '환상 동화'를 주제로 가진 도서에 이른다. '환상 동화' 주변의 연결을 옮기면 〈그림 3〉과 같다. 최상위의 '문학(예술)'은 사례에 서브그래프가 수집되지 않은 개체이며, 점선 상자로 표시하였다.



〈그림 3〉 주제 그래프의 연결 예시

주제 사이의 경로가 이어져 있어도 도서에 주제 연결이 없거나 사례를 정의한 분류와 같은 주제 하나만 있으면, 그 도서는 주제로 검색되지 않거나 다른 도서와 구분되지 않는다. 그 구성은 〈표 11〉과 같다. 주제가 부여되지 않은 도서가 5,331건이고, 부여된 주제가 사례를 정의한 분류기호와 같은 범주인 '한국 현대 소설' 하나뿐인 도서가 13,297건이다. 이 2가지 묶음을 합한 18,628건, 곧 사례 도서 53,548건의 34.79%는 주제로 서로 구분되지 않는다. '한국 현대 소설'은 주제 그래프에서 '한국 소설'과 '현대 소설'을 상위로 가지며, 사례 전체에서 도서 28,260건에 연결된 가장 큰 주제이다. 주제가 하나뿐인 그 밖의 도서 17,494건은 창작 그림책(8,501건)과 창작 동화(5,308건) 같은 아동서 계열이 대부분이고, 주제가 둘 이상인 도서 17,426건의 85.87%는 '한국 현대 소설'에 애정 소설이나 판타지 소설 같은 장르 주제를 더한 형태이다.

〈표 11〉 주제 연결의 구성

구분	도서(건)	비율
주제 연결 없음	5,331	9.96%
주제 하나 – '한국 현대 소설'	13,297	24.83%
주제 하나 – 그 밖의 주제	17,494	32.67%
주제 둘 이상	17,426	32.54%
전체	53,548	100%

3. 추론 과정에서의 적용

추론 과정에서 적용한 규칙은 OWL 2 RL(W3C, 2012)이며, 이는 OWL 2가 정한 선언 가운데 규칙의 형태로 실행할 수 있는 부분을 묶은 것이며, GraphDB 11.4.1의 내장 규칙 집합(ruleset) owl2-rl을 통해 구현하였다. RDFS의 클래스와 관계 계층은 자원의 상위 타입을 재진술하는 트리플을 도출하고, OWL의 역관계와 대칭 선언은 명시되지 않은 방향의 관계를 도출한다. 국가서지

온톨로지가 사용하는 선언이 역관계(owl:inverseOf), 동일성(owl:sameAs), 정의역과 치역처럼 OWL 2 RL이 다루는 범위 안에 있기 때문이다.

본 연구는 추론의 효과를 신규 질의 가능 트리플의 수로 확인한다. 신규 질의 가능 트리플은 적재된 트리플에는 없으나 선언에 따라 도출되어 질의로 확인할 수 있게 된 트리플을 말한다.

엔테일먼트 레짐을 적용하면 명시된 트리플과 별개로 도출되는 트리플이 7,576,641건에 이른다. 그러나 그중 97.60%는 이미 있는 타입의 상위 타입을 반복하는 rdf:type 트리플(5,872,637건)과 자원 자신과의 동일성(1,521,787건)을 나타내는 트리플이어서 탐색에 쓰이지 않는다. 나머지도 대부분 같은 연결을 상위 속성의 이름으로 재진술한 것이다(181,489건). 도출된 트리플의 구성은 〈표 12〉와 같다. 새로 질의할 수 있게 된 역방향 트리플은 728건이며, 모두 도서 사이의 축에서 나왔다. 이들은 번역과 후속, 보유판(補遺版, nlon:hasSupplement)처럼 역관계가 선언되어 있는 관계들이다. 유형별로는 번역 원작으로 향하는 관계가 406건, 번역본으로 향하는 관계가 232건, 후속저록이 81건, 선행저록이 7건, 보유판이 2건이다.

〈표 12〉 엔테일먼트 레짐 적용으로 도출된 트리플의 구성

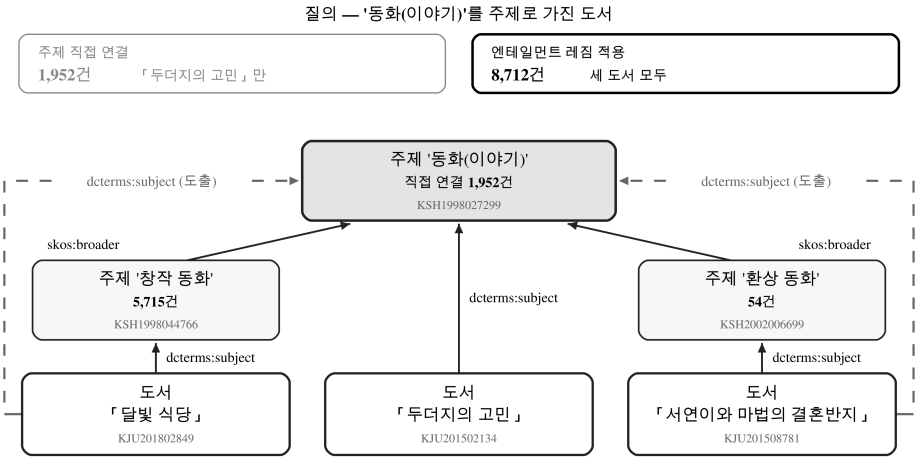
구분	트리플(건)
이미 있는 타입의 상위 타입 반복(rdf:type)	5,872,637
자원 자신과의 동일성(owl:sameAs)	1,521,787
같은 값을 상위 속성의 이름으로 반복	181,489
역방향 관계	728
합계	7,576,641

역관계 선언이 적용되면 양쪽이 모두 1,169건으로 일치하게 된다. 이 값은 관계의 상대편이 사례 밖 도서인 경우까지 포함하여 셈한 것으로, 양쪽이 모두 사례 안에서 타입이 확인되는 개체 사이만 세는 〈표 6〉의 도서와 도서 축과는 산정 범위가 다르다.

반면 도출이 일어나지 않은 곳은 선언이 없거나 데이터가 없는 곳이다. 규모가 가장 큰 저자 관계와 주제 관계는 역관계 선언이 없어 도출된 트리플이 한 건도 없다. 상위와 하위 총서 관계(nlon:subSeries와 nlon:mainSeries)는 반대로 역관계가 선언되어 있으나, 사례에 그 데이터가 없어 도출할 것이 없었다. owl:sameAs에서는 1,521,787건이 도출되었으나 전량이 각 자원을 그 자신과 같다고 적는 기계적 도출이다. 사례의 모든 자원에 owl:sameAs가 한 건씩 도출되었을 뿐, 서로 다른 식별자를 잇는 동일성은 한 건도 없다. 저자와 도서의 owl:sameAs 값은 문자열로 저장되어 있고, 주제에 URI로 저장된 815건은 상대가 사례 수집 범위 밖에 있기 때문이다. 클래스가 부여되지 않은 저자 391개는 추론을 거쳐도 저자 클래스(nlon:Author)로는 조회되지 않고, 더 넓은 사람 클래스(foaf:Person)로만 조회된다.

선언이 있을 때 무엇이 달라지는지는 SKOS의 이행 선언에서 확인된다. 사례의 도서 「두더지의

고민」은 주제 '동화(이야기)'를, 「달빛 식당」은 '창작 동화'를, 「서연이와 마법의 결혼반지」는 '환상 동화'와 각각 연결된다. 뒤의 주제 2개는 '동화(이야기)'를 상위 주제로 가지지만 도서 2권에 직접적으로 주제 속성에 '동화(이야기)'가 입력된 트리플은 없다. 그래서 '동화(이야기)'를 주제로 가진 도서를 묻는 질의를 수행하면 입력된 관계만 판단하여 1,952건의 도서, 예시 도서 3권 중에서는 「두더지의 고민」만 반환된다. SKOS는 상위 관계를 이행 관계의 하위 속성으로 선언하므로(W3C, 2009), 그 선언을 함께 적재하고 엔테일먼트 레짐을 적용하면 같은 질의가 하위 주제 78종을 거쳐 8,712건을 반환하며 도서 3권이 모두 그 안에 들어온다. 이 관계는 〈그림 4〉와 같다.



〈그림 4〉 SKOS 연결 관계를 통한 주제 확장 예시

앞 절에서 쓴 속성 경로로 하위 방향과 그 역방향을 함께 적용하면 같은 질의가 같은 8,712건을 반환한다. 곧 이 데이터에서 엔테일먼트 레짐이 더하는 것은 질의로 도달할 수 없던 범위가 아니라, 질의에 적어야 했던 경로를 선언이 대신 적는다는 것이다. 추론의 값은 도달 범위가 아니라 질의의 단순함에 있다.

4. 시맨틱 검색의 제한점 파악

본 절은 앞에서 확인한 탐색과 추론의 제한이 어떤 데이터 구조에서 비롯되는지를 3가지 문제로 정리한다.

첫째, 개체의 식별 요건에서 클래스와 그 개체가 갖추어지지 않았다. 출판사는 대응하는 클래스 선언 없이 문자열로만 존재하고, 분류와 총서는 클래스가 선언되어 있음에도 사례에 그 개체가 없어 선언된 구조를 채우지 못한다. 어느 쪽이든 문자열을 URI 형식으로 고쳐 되돌릴 수 있는 것이 아니라, 클래스를 선언하거나 문자열로 있는 값을 URI 개체로 세워 정의할 대상이다.

둘째, 의미 모델이 정의한 관계 속성의 종류가 부족하다. 도서와 저자를 잇는 관계는 dcterms:creator 하나여서, 지은이와 그린이, 옮긴이가 모두 같은 관계로 연결되고 역할을 구분해 따라가는 질의는 성립하지 않는다. 사례 안의 실질 연결도 도서와 주제, 도서와 저자, 주제와 주제, 도서와 도서의 클래스 쌍 4가지다.

셋째, 관계를 통한 연결 요건에서 이미 있는 개체와 관계가 관리되지 않아 연결이 끊기거나 한쪽만 채워진다. 이로 인하여 탐색에서 확장을 제한하는 관계 종류가 부족한 문제가 발생한다. 저자 13,198개는 타입 값이 문자열이어서 클래스 질의가 성립하지 않으며, URI 클래스를 함께 갖지 않은 391개는 어느 클래스로도 조회되지 않는다. 연결이 비어 있는 곳은 저자 URI와 연결되지 않은 도서 17,412건, 주제 URI가 연결되지 않은 도서 5,331건, 저자가 문자열로도 URI로도 없는 도서 527건이다. 연결이 있어도 변별을 주지 못하는 곳은 부여된 주제가 분류기호와 같은 범주인 '한국 현대 소설' 하나뿐인 도서 13,297건이며, 앞의 5,331건과 합하면 사례의 34.79%가 주제로 구분되지 않는다. 한 방향만 채워진 곳은 저자 관계 9,827건과 주제 관계 9,000건, 합하여 18,827건이고, 번역 관계는 2가지 방향의 수가 937건과 763건으로 어긋나 있다.

5. 데이터 구조 개선 방안

본 절은 앞 절에서 파악한 문제를 중심으로 데이터 구조의 개선 방안을 논의한다. 문자열로만 남아 있는 대상은 클래스가 없으면 온톨로지에 클래스를 선언하고, 그 값을 URI 개체로 세운다. 하나로 묶인 관계는 역할별로 나누어 세분화하고, 문자열로 남은 값과 비어 있는 연결은 정비한다. 이 3가지 조치를 클래스 신설, 관계의 세분화, 값과 연결의 정비로 부르며, 앞 절의 3가지 문제와 차례로 대응한다. 이 가운데 값과 연결의 정비는 역방향의 공백을 메우는 온톨로지 선언 보완을 포함한다. 클래스 신설은 대상에 따라 현재 데이터셋 안에서 실행할 수 있는 것과 원천 데이터의 입력까지 거슬러 올라가는 것으로 나뉜다. 조치별 개선 후보를 정리하면 〈표 13〉과 같다. 〈표 13〉의 저자 직업과 활동 분야 값은 저자 URI 13,198개를 기준으로 한 비율이며, 직업은 schema:jobTitle, 활동 분야는 nlon:fieldOfActivity 값이다.

〈표 13〉 개선 조치별 후보

갈래	후보	규모와 상태	비고
클래스 신설	출판사	문자열 60,307건 (고유 6,123종)	클래스 선언 없음
	총서	총서명 32,331건 (고유 5,084종)	클래스 선언은 있으나 개체 없음
	저자의 직업	저자 98.80% 보유 (고유 709종)	종류가 한정된 문자열
	저자의 활동 분야	저자 전원 보유 (고유 620종)	종류가 한정된 문자열
	수상 정보	주기사항 1,486건 (도서 1,452건)	표기가 일관되지 못하고 표제와 주기에 분산
관계의 세분화	저자 역할 세분	dcterms:creator 단일	직업 값이 역할 판별의 단서
값과 연결의 정비	연관저록 연결	이판저록 256건, 선행저록 127건, 후속저록 53건(URI)	URI 연결 밖은 표제 문자열

가. 클래스 신설

값 자체는 데이터에 들어 있으나 문자열로 저장되어 연결로 기능하지 못하는 값을 URI 개체로 만들면 같은 값을 가진 도서가 그 개체를 통해 묶이고, 이를 경유하는 탐색이 성립한다.

출판사는 한 출판사가 여러 도서에 걸쳐 나타나는데도 대응하는 클래스 선언 없이 문자열로만 존재한다. 다만 같은 출판사가 여러 표기로 저장되어 있어, 개체를 세우려면 표기를 병합하는 판별이 앞선다.

총서는 클래스가 선언되어 있으나 사례에서 개체로 나타나지 않는다. 총서를 잇는 관계로 이어진 도서가 24건 있으나 그 URI를 주체로 하는 속성이 데이터에 없어, 분류와 마찬가지로 선언된 구조가 데이터에서 비어 있다. 총서 개체를 세우려면 출판사 개체의 판별이 먼저 이루어져야 한다. 흔한 총서명을 여러 출판사가 함께 쓰고 있기 때문이다. 총서 안의 순서를 잇는 권차도 여러 방식으로 저장되어 있다.

저자의 직업과 활동 분야는 저자들이 나누어 가지는 값이면서 종류가 한정되어 있어, 개체나 관계로 승격하면 이를 조건으로 저자를 가르는 질의가 성립한다. 다만 표기만 다른 값은 정리로 합쳐지지만, '일러스트레이터'와 '삽화가', '그림작가'처럼 가까운 일이 외래어와 우리말로 갈린 값은 표기 정리로 합쳐지지 않는다.

이 후보들의 개체를 세우기 전에 거쳐야 하는 표기 정비 과제는 〈표 14〉와 같다.

〈표 14〉 클래스 신설 후보의 표기 정비 과제와 예시

후보	정비 과제	예시
출판사	접미어의 유무	'사계절' 116건, '사계절출판사' 104건
	띄어쓰기	'어울림 출판사' 149건, '어울림출판사' 110건, '어울림' 47건
	한글과 영문의 병기 순서	'Bookk(부크크)' 417건, '부크크(Bookk)' 154건
총서	별개의 출판사가 같은 이름의 총서를 냄	'가을'이 출판사 12종과 결합
	같은 총서의 표기가 나뉨	'즐거운 동화 여행' 155건, '즐거운 동화여행' 26건
	권차 표기의 비통일	'1권', '제1권', '제1집', 'No.01', '第1輯', '두번째'
저자 직업	표기 방식만 다른 값	'그림책작가' 440건, '그림책 작가' 71건
	표기 방식만 다른 값	'수필가' 154건, '수필가[隨筆家]' 95건
	외래어와 우리말로 나뉜 값	'일러스트레이터' 2,005건, '삽화가' 23건, '그림작가' 34건

수상 정보는 값이 데이터 안에 흩어져 있다는 점이 다르다. 수상 주기(nlon:awardsNote)가 도서 1,452건에 있으나 표기가 일정하지 않다. 같은 상이 한글과 한자로 나뉘어 저장되고, 표제관련정보(nlon:remainderOfTitle) 212건에도 흩어져 있어, 개체로 세우려면 표기를 정리하는 재입력이 필요하다. 수상 정보가 URI 개체가 되면 상에서 수상 도서로, 다시 그 저자로 이어지는 질의가 성립한다.

나. 관계의 세분화

앞에서 정리한 둘째 문제가 여기에 해당한다. 관계 세분화는 dcterms:creator 하나로 묶인 도서

와 저자의 연결을 역할별 관계로 나누는 방향이다. 그림과 번역, 편집처럼 역할이 다른 저자가 모두 동등한 저자로 연결되어 있어, 역할을 구분한 관계가 도입되면 지은 책과 옮긴 책, 그린 책을 갈라 따라가는 질의가 성립한다. 판별의 단서는 저자의 직업 값에 이미 있으나, 직업은 저자에 붙는 값이고 역할은 도서와 저자 사이의 관계여서 한 저자가 어떤 책은 짓고 다른 책은 옮기는 경우는 가려지지 않는다. 역할을 구분하는 관계의 설계와 기존 연결의 재기술이 따른다.

연관저록의 연결은 저자 역할과 사정이 다르다. 도서와 도서를 잇는 관계 어휘는 유형별로 온톨로지에 이미 선언되어 있다. 문제는 그 관계가 실제로 거의 입력되지 않았다는 것이다. 사례 도서 53,548건 가운데 연관저록 관계를 하나라도 가진 도서는 1,573건이다. 이 가운데 번역 관계를 가진 도서가 1,187건으로 대부분이고, 이판저록이 253건, 선행저록과 후속저록이 133건이며, 나머지 유형은 소수다. 따라서 연관저록에 필요한 것은 관계를 새로 나누는 일이 아니라, 이미 관계 속성으로 선언된 관계를 실제로 채우는 일이다.

다. 값과 연결의 정비와 선언 보완

앞에서 정리한 셋째 문제가 여기에 해당한다. 값과 연결의 정비는 문자열로 남은 타입 값을 형식에 맞게 고치고, 저자나 주제와 연결되지 않은 도서에 연결을 채우는 방향이다. 연결이 있어도 변별을 주지 못하는 곳 역시 대상이다. 가장 넓은 주제인 '한국 현대 소설'은 도서 28,260건에 부여되어 하나의 보편적인 연결에 서로 다른 대상이 묶여 있는데, 그 가운데 14,963건에는 애정 소설이나 판타지 소설 같은 장르 주제가 함께 부여되어 있다. 정비의 방향은 주제를 새로 만드는 것이 아니라 이미 부여된 주제를 상하위로 잇고 주제가 없는 도서에 기술을 채우는 것이다.

온톨로지 선언 보완의 효과는 제안한 선언을 추가한 데이터에서 측정하였다. 저자 관계와 주제 관계에 역관계를 선언하자 역방향 결손 트리플 18,827건이 전량 도출되어 질의로 확인할 수 있게 되었다. 이로써 도서 쪽에 적힌 연결의 80.46%와 86.54%만을 반환하던 저자에서 도서로, 주제에서 도서로 가는 질의가 전량을 반환하게 된다. 여기에 SKOS가 대칭으로 규정한 관련 관계의 대칭 선언을 더하자 트리플 7,042건이 추가로 도출되었다. 관련 관계는 사례에서 7,512건 가운데 470건만 역방향이 함께 있었으나, 대칭 선언을 적용한 뒤에는 전체가 양방향으로 확인된다. 이 3가지 선언만으로 새로 질의할 수 있게 된 트리플은 25,869건이며, 이 과정에서 새로 삽입한 트리플은 없다. 출판사를 개체로 정리하면 같은 출판사의 다른 도서를 관계로 찾는 질의를 수행할 수 있다. 저자 관계를 역할에 따라 구분하면, 직업 속성(schema:jobTitle) 값이 '일러스트레이터'나 '그림작가'처럼 그림에 해당하는 저자 URI 2,488개와 번역에 해당하는 376개, 편집에 해당하는 193개를 역할로 구분할 수 있고, 질의를 세분화하여 수행할 수 있다. 앞 절에서 확인한 SKOS 이행 선언은 데이터를 고치는 것이 아니라 데이터가 이미 쓰고 있는 표준 어휘의 정의를 함께 싣는 것이므로 개선 조치로 세지 않았다. 개별 개체에서도 공백이 확인된다. 예를 들어 연구에서

사용한 한국 현대 소설 사례 범위 안에서 저자 백희나가 dcterms:creator로 연결된 도서는 113건이고, 저자 백희나를 주체로 nlon:create로 연결된 도서는 91건인데, 중복을 제외하면 합계 159건이다. 159건 가운데 46건은 도서 개체에서 백희나를 찾아갈 수 없고, 68건은 백희나의 nlon:create로 탐색할 수 없는 것이다. 역관계를 선언하면 이 68건이 전량 도출되어 159건이 반환될 수 있다.

Ⅴ. 논의 및 결론

본 연구는 국가서지 LOD가 다운로드 형식으로 제공하는 데이터셋을 트리플 저장소에 적재하여, 데이터 구조의 특징을 관찰하고 시맨틱 검색의 수행 가능성을 파악하였다.

국가서지 LOD 데이터셋의 사례에서 확인한 사실은 다음과 같다. 개체의 식별은 부분적으로만 갖추어져 있다. 도서와 저자, 주제는 URI 개체로 제공된다. 저자는 도서에 이름 문자열과 URI 연결로 나뉘어 표현되는데, 이름 문자열을 가진 도서 52,840건 가운데 32.95%는 저자 URI 연결이 없어 문자열로만 남는다. 출판사는 개체로 서술되어 있지 않다. 의미 모델을 정의한 온톨로지에서는 선언과 데이터가 어긋난다. 온톨로지가 선언한 클래스 31종 가운데 사례에 타입 값으로 나타나는 것은 6종이고, 분류와 총서는 선언된 구조를 데이터가 채우지 못하며 출판사는 선언 자체가 없다. 관계를 통한 연결은 도서와 주제, 도서와 저자, 주제와 주제, 도서와 도서의 4가지에서 성립하였으나, 규모가 큰 저자 관계와 주제 관계에서 역방향 트리플 18,827건이 비어 있다. 활용의 요건인 시맨틱 검색은 명시된 관계의 범위 안에서 수행되었다. 그래프 패턴 질의는 관계를 한 단계 따라가는 확인과 집계, 2단계의 확장까지 성립하였고, 속성 경로를 통한 주제 확장도 성립하였으나 상위와 하위 방향의 확장은 4단계에서 더 이상 새로운 주제에 도달하지 않았다. 온톨로지에 이미 들어 있는 역관계 선언은 번역과 후속, 보유판처럼 트리플 수가 적은 관계에만 걸려 있어 추론이 더하는 트리플은 728건뿐이었으나, 저자 관계와 주제 관계에 역관계를, 관련 관계에 대칭을 선언하는 3가지를 추가하면 트리플 25,869건을 새로 질의할 수 있었다. 이 제한을 낳는 구조는 클래스와 개체의 부족, 관계 종류의 부족, 값과 연결의 관리 공백이라는 3가지 문제로 정리된다.

이 결과를 토대로 데이터 구조의 개선 방안을 다음과 같이 제안하였다. 첫째는 출판사처럼 문자열로만 남아 있는 대상에 클래스를 신설하고 그 값을 URI 개체로 세우는 것이다. 후보로 출판사와 총서, 저자의 직업과 활동 분야, 수상 정보를 제시하였다. 둘째는 도서와 저자를 잇는 관계를 역할별로 세분하고 문자열로 남은 연관저록을 선언된 관계 어휘로 잇는 것이다. 셋째는 문자열로 남은 값과 비어 있는 연결을 정비하고 역방향 관계와 관련 관계의 대칭을 온톨로지 선언으로 보완하는 것이다.

본 연구는 국가서지 LOD의 사례에서 확인한 제한을 3가지 문제로 정리하고, 서지 구조 전환 과정에서 활용할 수 있는 개선 방안을 제언하였다는 점에 의의가 있다. 본 연구의 한계는 각각

후속 연구의 과제로 이어진다. 확인은 질의 언어로 관계를 어디까지 따라갈 수 있는지에 한정하였다. SPARQL로 성립한 탐색과 추론을 이용자가 쓰는 검색 화면에 올리려면 그 사이에 어떤 계층이 필요한지는 구현과 실증을 함께 요구하는 문제로 남는다. 사례는 하나의 분류 계열에 한정하였고, 그에 따라 주제 그래프의 규모도 사례 안으로 제한되었다. 다른 분야로 확장하면 문제의 구조가 어떻게 달라지는지 확인할 필요가 있다.

참 고 문 헌

국가도서관위원회 (2024). 제4차 도서관발전종합계획(2024~2028).

국가도서관위원회 (2026). 제4차 도서관발전종합계획 2026년도 시행계획.

국립중앙도서관 (2013). 국가서지의 Linked Data 변환 및 응용서비스 개발 완료보고서.

국립중앙도서관 (2024). 국가서지에 대한 차세대 서지 구조 적용방안 연구 (11-1371029-000260-01).

국립중앙도서관 (발행년불명). RDF 매핑. 국가서지 링크드 오픈 데이터. 출처: https://lod.nl.go.kr/home/service/rdf.do

노지현 (2015). 주제명 데이터를 통해 본 현행 목록의 품질과 과제. 한국도서관·정보학회지, 46(4), 379-402. https://doi.org/10.16981/kliss.46.4.201512.379

노지현 (2019). 편목의 관점에서 본 링크드 데이터: 현황과 과제. 한국도서관·정보학회지, 50(3), 71-95. https://doi.org/10.16981/kliss.50.3.201909.71

도서관법. 법률 제21090호.

박진호 (2026). SHACL과 ShEx를 활용한 국가서지LOD의 품질 평가: 온톨로지 설계 검증과 실데이터 검증의 이중 층위 분석. 한국비블리아학회지, 37(2), 321-347. https://doi.org/10.14699/kbiblia.2026.37.2.321

박진호, 이미화, 송민선 (2025). BIBFRAME 도입을 위한 전담조직 운영 컨설팅 (11-1371029-100018-01). 국립중앙도서관.

백지원, 정연경 (2014). 국립중앙도서관 주제명표목표 검색 시스템 개선 방안에 관한 연구. 정보관리학회지, 31(1), 31-51. https://doi.org/10.3743/KOSIM.2014.31.1.031

이미화, 이지원 (2021). 주요국 국가서지 현황조사를 통한 국가서지의 최신 경향 분석. 한국비블리아학회지, 32(1), 35-57. https://doi.org/10.14699/kbiblia.2021.32.1.035

이성숙 (2020). 국내 도서관 링크드 오픈 데이터 구축과 발행의 개선방안 연구. 정보관리학회지, 37(2), 145-169. https://doi.org/10.3743/KOSIM.2020.37.2.145

이혜경, 이용구 (2023). 주제명 활용 분석을 통한 국립중앙도서관 주제명표목표의 현황 연구.

정보관리학회지, 40(2), 157-182. https://doi.org/10.3743/KOSIM.2023.40.2.157

정도헌 (2024). 시맨틱 추론 규칙을 이용한 대규모 언어 자원의 품질 고도화 방안: 국립중앙도서관 주제명표목표를 중심으로. 한국도서관·정보학회지, 55(4), 245-263. https://doi.org/10.16981/kliss.55.4.202412.245

조명대 (2010). 도서관에서의 Linked Data 활용방안에 관한 연구. 한국문헌정보학회지, 44(1), 181-198. https://doi.org/10.4275/KSLIS.2010.44.1.181

최윤경, 정연경 (2014). 국립중앙도서관 주제명표목표의 고품질화 방안에 관한 연구. 한국문헌정보학회지, 48(1), 75-95. https://doi.org/10.4275/KSLIS.2014.48.1.075

Aalberg, T., Tallerås, K., & Massey, D. (2019). The impact of new bibliographic models on the search experience. Information Research, 24(4), paper colis1915. https://informationr.net/ir/24-4/colis/colis1915.html

Balog, K. (2018). Entity-Oriented Search. The Information Retrieval Series, 39. Cham: Springer.

Bast, H., Buchhold, B., & Haussmann, E. (2016). Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval, 10(2-3), 119-271. https://doi.org/10.1561/1500000032

Berners-Lee, T. (2006). Linked Data. Design Issues. Available: https://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - the story so far. International Journal on Semantic Web and Information Systems, 5(3), 1-22. https://doi.org/10.4018/jswis.2009081901

Deliot, C., Wilson, N., Costabello, L., & Vandenbussche, P.-Y. (2016). The British National Bibliography: Who uses our linked data? Proceedings of the International Conference on Dublin Core and Metadata Applications 2016, 24-33. https://doi.org/10.23106/dcmi.952137546

Diao, J. (2018). Evolution of Western library catalogs: The rising expectations of users. International Journal of Librarianship, 3(1), 78-93. https://doi.org/10.23974/ijol.2018.vol3.1.69

Gnoli, C., Golub, K., Haynes, D., Salaba, A., Shiri, A., & Slavic, A. (2024). Library catalog's search interface: Making the most of subject metadata. Knowledge Organization, 51(3), 169-186. https://doi.org/10.5771/0943-7444-2024-3-169

Guha, R., McCool, R., & Miller, E. (2003). Semantic search. Proceedings of the 12th International Conference on World Wide Web, 700-709. https://doi.org/10.1145/775152.775250

World Wide Web Consortium (2009). SKOS Simple Knowledge Organization System Reference. W3C Recommendation. Available: https://www.w3.org/TR/skos-reference/

World Wide Web Consortium (2011). Library Linked Data Incubator Group Final Report. W3C Incubator Group Report. Available: https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

World Wide Web Consortium (2012). OWL 2 Web Ontology Language Profiles (Second Edition). W3C Recommendation. Available: https://www.w3.org/TR/owl2-profiles/

World Wide Web Consortium (2013a). SPARQL 1.1 Entailment Regimes. W3C Recommendation. Available: https://www.w3.org/TR/sparql11-entailment/

World Wide Web Consortium (2013b). SPARQL 1.1 Query Language. W3C Recommendation. Available: https://www.w3.org/TR/sparql11-query/

World Wide Web Consortium (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. Available: https://www.w3.org/TR/rdf11-concepts/

• 국한문 참고문헌의 영문 표기
(English translation / Romanization of references originally written in Korean)

Baek, Ji-Won & Chung, Yeon Kyoung (2014). A study on improving access & retrieval system of the National Library of Korea subject headings. Journal of the Korean Society for Information Management, 31(1), 31-51. https://doi.org/10.3743/KOSIM.2014.31.1.031

Cho, Myungdae (2010). A study on applications for linked data in libraries. Journal of the Korean Society for Library and Information Science, 44(1), 181-198. https://doi.org/10.4275/KSLIS.2010.44.1.181

Choi, Yoon Kyung & Chung, Yeon Kyoung (2014). A study on improvements for high quality in National Library of Korea subject headings list. Journal of the Korean Society for Library and Information Science, 48(1), 75-95. https://doi.org/10.4275/KSLIS.2014.48.1.075

Jeong, Do-heon (2024). Method for enhancing the quality of large-scale lexical resources using semantic inference rules: Focusing on the National Library of Korea subject headings. Journal of Korean Library and Information Science Society, 55(4), 245-263. https://doi.org/10.16981/kliss.55.4.202412.245

Lee, Hyekyung & Lee, Yong-Gu (2023). A study on the current status of National Library of Korea subject headings list through utilization analysis of subject headings. Journal

of the Korean Society for Information Management, 40(2), 157-182. https://doi.org/10.3743/KOSIM.2023.40.2.157

Lee, Mihwa & Lee, Ji Won (2021). Current trends for national bibliography through analyzing the status of representative national bibliographies. Journal of the Korean Biblia Society for Library and Information Science, 32(1), 35-57. https://doi.org/10.14699/kbiblia.2021.32.1.035

Lee, Sung Sook (2020). A study on the improvement plan of publishing in library linked open data. Journal of the Korean Society for Information Management, 37(2), 145-169. https://doi.org/10.3743/KOSIM.2020.37.2.145

Libraries Act. Act No. 21090.

National Library Committee (2024). The 4th Comprehensive Plan for Library Development (2024-2028).

National Library Committee (2026). The 4th Comprehensive Plan for Library Development: 2026 Annual Implementation Plan.

National Library of Korea (2013). Final Report on the Linked Data Conversion of the National Bibliography and the Development of Application Services.

National Library of Korea (2024). A Study on the Application of Next-Generation Bibliographic Structures to the National Bibliography (11-1371029-000260-01).

National Library of Korea (n.d.). RDF mapping. National Bibliography Linked Open Data. Available: https://lod.nl.go.kr/home/service/rdf.do

Park, Jin Ho (2026). A quality assessment of the national bibliographic LOD using SHACL and ShEx: A dual-layer analysis of ontology design and instance data validation. Journal of the Korean Biblia Society for Library and Information Science, 37(2), 321-347. https://doi.org/10.14699/kbiblia.2026.37.2.321

Park, Jin Ho, Lee, Mihwa, & Song, Min-sun (2025). Consulting on the Operation of a Dedicated Organization for BIBFRAME Adoption (11-1371029-100018-01). National Library of Korea.

Rho, Jee-Hyun (2015). A study on the quality of subject data in library catalogs. Journal of Korean Library and Information Science Society, 46(4), 379-402. https://doi.org/10.16981/kliss.46.4.201512.379

Rho, Jee-Hyun (2019). The current state and challenges of linked data in library cataloging. Journal of Korean Library and Information Science Society, 50(3), 71-95. https://doi.org/10.16981/kliss.50.3.201909.71